7-24-2017 12:00 AM

Software-defined Networking enabled Resource Management and Security Provisioning in 5G Heterogeneous Networks

Xiaoyu Duan
The University of Western Ontario

Supervisor
Dr. Xianbin Wang
The University of Western Ontario

Graduate Program in Electrical and Computer Engineering
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Abstract

Due to the explosive growth of mobile data traffic and the shortage of spectral resources, 5G networks are envisioned to have a densified heterogeneous network (HetNet) architecture, combining multiple radio access technologies (multi-RATs) into a single holistic network. The co-existing of multi-tier architectures bring new challenges, especially on resource management and security provisioning, due to the lack of common interface and consistent policy across HetNets. In this thesis, we aim to address the technical challenges of data traffic management, coordinated spectrum sharing and security provisioning in 5G HetNets through the introduction of a programmable management platform based on Software-defined networking (SDN).

To address the spectrum shortage problem in cellular networks, cellular data traffic is efficiently offloaded to the Wi-Fi network, and the quality of service of user applications is guaranteed with the proposed delay tolerance based partial data offloading algorithm. A two-layered information collection is also applied to best load balancing decision-making. Numerical results show that the proposed schemes exploit an SDN controller's global view of the HetNets and take optimized resource allocation decisions. To support growing vehicle-generated data traffic in 5G-vehicle ad hoc networks (VANET), SDN-enabled adaptive vehicle clustering algorithm is proposed based on the real-time road traffic condition collected from HetNet infrastructure. Traffic offloading is achieved within each cluster and dynamic beamformed transmission is also applied to improve trunk link communication quality.

To further achieve a coordinated spectrum sharing across HetNets, an SDN enabled orchestrated spectrum sharing scheme that integrates participating HetNets into an amalgamated network through a common configuration interface and real-time information exchange is proposed. In order to effectively protect incumbent users, a real-time 3D interference map is developed to guide the spectrum access based on the SDN global view. MATLAB simulations confirm that average interference at incumbents is reduced as well as the average number of denied access.

Moreover, to tackle the contradiction between more stringent latency requirement of 5G and the potential delay induced by frequent authentications in 5G small cells and HetNets, an SDN-enabled fast authentication scheme is proposed in this thesis to simplify authentication

handover, through sharing of user-dependent secure context information (SCI) among related access points. The proposed SCI is a weighted combination of user-specific attributes, which provides unique fingerprint of the specific device without additional hardware and computation cost. Numerical results show that the proposed non-cryptographic authentication scheme achieves comparable security with traditional cryptographic algorithms, while reduces authentication complexity and latency especially when network load is high.

Keywords: 5G HetNets, authentication, data offloading, spectrum sharing, VANET

Dedication

To my parents and husband

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Dr. Xianbin Wang, for his guidance, patience and encouragement in developing my research. It was his enlightening supervisions that inspired me to explore novel research areas and broadened my views in the research area.

Sincere thanks to Professor Luiz Fernando Capretz, Raveendra Rao, Xingfu Zou from Western University, and Professor Jahangir Hossain from the University of British Columbia for serving as my thesis examiners and a critical reading of the dissertation. Their insightful advice and comments improves the quality of this dissertation.

Thanks go to all present and former members of our research group for the time that we spent both at work and after work. I would give my best regards for their success in both study and life. I would like to extend my thanks to all my friends at UWO. The last four years have been full of fun and warmth because of your accompany and friendship. As always, I wish to thank my husband and my parents for their love, support and encouragement throughout these years.

Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	Third Generation partnership project
4G	Fourth generation mobile communication system
5G	Fifth generation mobile communication system
AAA	Authentication, authorization, and accounting
AMC	Adaptive modulation and coding
AOA	Angle of arrival
AP	Access point
AWGN	Additive white gaussian noise
BER	Bit error rate
BS	Base station
CA	Carrier aggregation
CDF	Cumulative distribution function
CFO	Carrier frequency offset
CH	Cluster head
CIR	Channel impulse response
CoMP	Coordinated multipoint
CSI	Channel state information
D2D	Device-to-device
DSRC	Dedicated short range communications
eICIC	Enhanced intercell interference coordination
eMBB	Enhanced mobile broadband
XML	Extensible markup language
FCD	Floating car data
GDB	Global database

HetNet Heterogeneous network

I/Q In-phase/quadrature

IVD Inter-vehicular distance

GPS Global positioning system

LB Load balancing

LBT Last beacon time

LDB Local database

LDPC Low-density parity-check

LRT Likelihood ratio test

LTE Long-term evolution

LTE-A LTE-Advanced

MAC Media access control

MANET Mobile ad hoc network

MIMO Multiple input multiple output

mMTC Massive machine type communications

MMSE Minimum mean square error

MRT Maximum ratio transmission

MTC Machine type communications

NetConf Network configuration protocol

NFV Network functions virtualization

NMSE Normalized mean square error

NOMM Non-orthogonal multiplexed modulation

ONF Open network foundation

PDF Probability density function

PIT Predicted inhabitant time

QAM Quadrature amplitude modulation

QoS Quality of service

QPSK	Quadrature phase shift keying
RATs	Radio access technologies
RAN	Radio access network
RSSI	Received signal strength indicator
RSU	Road side units
SCI	Secure-context-information
SDN	Software-defined networking
SER	Symbol error rate
SINR	Signal to interference and noise ratio
SNMP	Simple network management protocol
TFT	Traffic flow template
TCP	Transmission control protocol
UE	User equipment
URLLC	Ultra-reliable and low-latency communications
V2I	Vehicle-to-Infrastructure
VANET	Vehicular ad hoc network
Wi-Fi	Wireless fidelity
WAVE	Wireless access in vehicular environments
WER	Word error rate
ZF	Zero Forcing

Chapter 1

Introduction

1.1 Evolution of Wireless Communication Systems

Over the last few decades, global communication technologies have experienced rapid development and have become indispensable to modern society, especially in the field of wireless communications. From the 1980's to the present, mobile services have evolved from 2G to 3G, Long-term Evolution (LTE), LTE-Advanced (LTE-A) with significantly increased coverage, speed and flexibility, and have also surpassed that of fixed communication technologies. With the advent of smart and media-rich mobile devices, cellular networks have witnessed an unprecedented growth in mobile data traffic due to ubiquitous mobile Internet access, traffic-intensive social applications, and cloud-based services [3]. Moreover, next generation cellular networks are expected to support a broad range of applications with different operational requirements. For example, automation, live video and inter-gaming are delay sensitive, while some other applications demand numerous connections, such as smart cities or Internet of things (IOT) [4]. 5G also needs to push the envelope of performance to provide higher capacity, lower latency, high reliability and greater mobility support. Specifically, emerging applications, such as virtual reality, industrial control, traffic safety, and mobile video are expected to become the mainstream in 5G [5]. Hence, the awareness of user application requirements and real-time information exchange is crucial for future cellular networks to support the large traffic volume and meet the Quality of Service (QoS) demands of diverse user applications.

1.1.1 Increasing Gap between Growing Data Traffic and Slow Upgrading Wireless Technologies

According to Cisco's networking visual index report published in February 2017 [6], data traffic grew 63 % in 2016 and reached 7.2 exabytes per month at the end of 2016 compared with 4.4 exabytes per month at the end of 2015. Global mobile data traffic is expected to grow at a compound annual rate of 47 % by 2021, reaching 49.0 exabytes per month in 2021. The huge demand for supporting the increasing data traffic has caused densified base station (BS)/ access point (AP) deployments. Deploying more BSs/ APs is one of the most straightforward ways to serve more users with increased data rate and coverage, and the other methods include wider spectrum using millimeter-wave transmission or greater spatial diversity of massive multiple-input, multiple-output (MIMO).

However, there is a gap between the growing demands on the cellular networks and the growth rate of wireless technologies. The mobile data traffic is expected to outgrow the capabilities of current fourth generation (4G) and Long Term Evolution (LTE) infrastructures by 2020. Current cellular networks are not capable of sustaining such high traffic volumes due to the slow updating of infrastructures and the shortage of spectrum resources. The relatively narrow usable frequency bands between several hundred MHz and a few GHz have been almost entirely occupied by a variety of licensed or unlicensed networks including 2G, 3G, LTE, LTE-A, and Wi-Fi.

1.1.2 Provisioning of Different QoS Requirement from Diverse User Applications

Back to ten years ago, voice and text service accounts for more than 80 percent of wireless communication markets and smartphones were still new concept gradually entering people's life. Given the unimaginably rapid emergence of new technologies in the wireless communication industry, mobile data traffic has grown 18-fold over the past five years, and mobile video traffic accounted for 60 percent of total mobile data traffic in 2016 [6]. Among the dramatically increased data traffic and diverse traffic types, different applications need different sets of services: a multimedia flow may prefer timeliness to reliable delivery, while IP telephony can

be tolerant to packet loss, or in some cases, to bit errors.

As such, QoS provisioning becomes essential in wireless communication to allocate resources efficiently and improve the user experience. Different applications have different operational constraints, for example, online gaming or multimedia are bandwidth aggressive, while real-time video calls are delay sensitive [4]. Enterprise level communication has higher latency and security constraints. Cellular networks are thus expected to support diverse types of applications with different quality demands and also offer ubiquitous and global connectivity for everything (users, devices, sensors, machines). Therefore, the next generation 5G network functionalities should be adaptable to the varying channel conditions, traffic load, as well as user application requirements.

1.2 Challenges of Current Wireless Communication Networks

1.2.1 Heterogeneity of Cellular Network Infrastructures

With the dramatically increased mobile and smart device numbers and their strong demands for bandwidth-hungry applications (e.g., mobile TV, inter-gaming, and live video), high data rate and low latency wireless services are in urgent need of the mobile network operators. In respect to this, 5G communication standards should include advanced technologies, such as enhanced multiple-input, multiple-output (MIMO) transmission (up to eight antenna pairs), Coordinated Multi-point transmission and reception (CoMP) and carrier aggregation (CA), to meet these requirements. CA can increase the bandwidth that can be allocated to end users through the concurrent utilization of different frequency carriers, while MIMO enhances the multi-antenna techniques with up to 88 antenna arrays. With CoMP transmission and reception, service outage probability at cell edge is envisioned to be dramatically decreased as multiple neighbor cells can coordinate their scheduling or transmission to serve edge users altogether.

However, due to the fact that spectral efficiency is approaching theoretical limits [7], all the above techniques may not always guarantee significant enhancement and meet the growing data traffic requirement. For example, the received signal powers are low after attenuation in, e.g., indoor scenarios (residential or office) under low signal-to-interference-plus-noise ratio

(SINR) conditions. Therefore, increasing node deployment density has been widely agreed as a possible solution to provide ubiquitous coverage and larger data rate.

Moreover, as the frequency bands between several hundred MHz and a few GHz that are most suitable for wireless communication have been almost entirely occupied by a variety of licensed or unlicensed networks including 2G, 3G, LTE, LTE-A and Wi-Fi, academia and industry have been exploring new bandwidths and idle spectra in the millimeter wave (mmW) range of $30 \sim 300GHz$. Millimeter wave, which is also known as extremely high frequency (EHF) or very high frequency (VHF), is an undeveloped band that can be used for high-speed wireless communications [8]. However, due to the reduced signal propagation characteristics at extremely high frequencies, mmWs have high atmospheric attenuation, and even rain and humidity can impact performance or reduce signal strength. Cell size thus has to be largely reduced in mmW bands.

Hence, the fifth generation (5G) networks are expected to have a densified heterogeneous network (HetNet) architecture, which combines multiple radio access technologies (multi-RATs), such as small cells and mmW communications into a single holistic network [9]. Both academia and industry believe that 5G networks will be heterogeneous with small cell deployment and overlay coverage, as shown in Fig. 1.1. Cellular networks operating at low frequencies could provide wide area coverage, mobility support and control, while small cells operating at higher frequencies guarantee high data rates given their spectral and energy efficiency.

1.2.2 Increased Management Complexity of HetNets

Although the heterogeneity of cellular networks is a natural evolution and is beneficial in boosting data rate and capacity, there are still challenges to be solved. Network densification using low-power small cells is widely considered to be a critical element towards low cost, high capacity 5G communications. However, the massive deployment of small cells poses potential challenges in network management, including interference alignment, extensive back-hauling, and inconsistent resource allocation/security mechanisms over HetNet. Network administration and service provisioning are challenging in this multi-tier model due to the increased

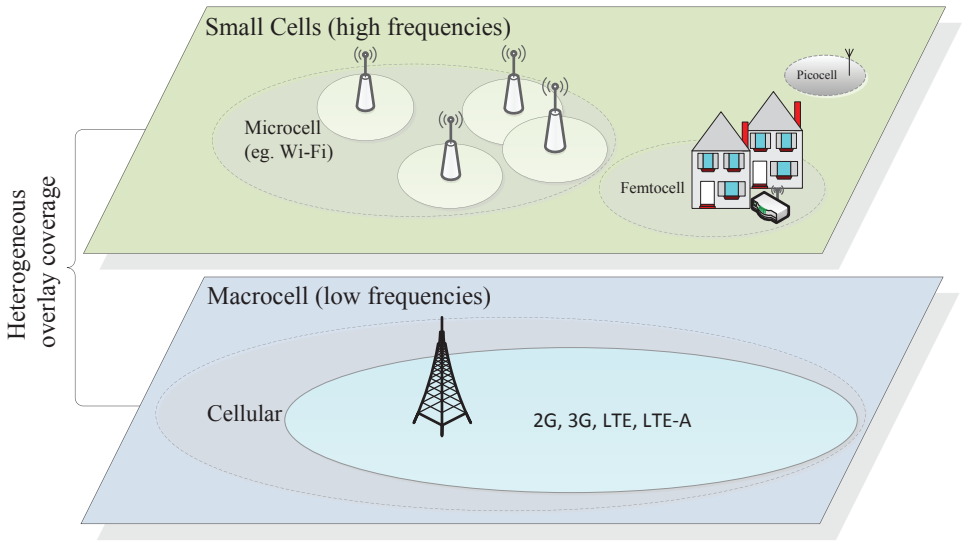

Figure 1.1: Fifth generation heterogeneous network structure with densified small cells and overlay coverage.

number of base-stations, the complexity of network architecture, and the vendor-specific equipment.

On the one hand, the limited information exchange among heterogeneous network infrastructures adds to the complexity of network management, which is further compounded by the vendor-specific equipment relying on different configuration interfaces. The ossification of the network infrastructure complicates protocol upgrades. That is to say, to make a new protocol available to applications, it has to be upgraded at all the paths traversing the network. The innovation and flexibility of the network are stifled due to the inconvenience in network adaptation and expansion. Moreover, operators only have indirect control of the network operation and resources, and there is hardly or no effective timely information sharing among the independent, heterogeneous access networks for coordination. For example, the layered HetNets architecture requires flexible deployment and high-level monitoring to avoid coverage blind spots. With smaller cell size due to the use of mmW bands, frequent handover and authentications for users moving across HetNets also need to be handled with more efficient security provisioning mechanisms.

Moreover, cellular networks have been preserving an application agnostic and base station (BS) centric architecture historically. The network functions or policies are performed with no consideration of specific user applications or relative locations, causing inefficient resource

management in both access networks and core networks. 5G networks are expected to support 1000-fold capacity increases with at least 100-fold in the peak data rate to reach 10Gbps and one magnitude smaller delay constraint from 4G [4]. Ultra-reliable and low-latency communications (URLLC) and Internet of Things with potential numbers of various connectible devices are also to be supported. All in all, 5G should be cost-efficient, flexibly deployable, elastic, and above all programmable. In order to obtain the desired 5G performance, a flexible HetNets management and QoS-aware resource allocation have become necessities. However, the traditional usage of vendor-specific networking equipment can neither dynamically scale with mobile traffic nor be easily upgraded with new functions.

Moreover, 5G users may leave one cell and join another more frequently with reduced cell size, which could introduce an excessive handover-induced latency in 5G. Therefore, new technologies are needed to provide intelligent control over the 5G HetNet for consistent and efficient traffic management as well as security management.

It has been widely accepted worldwide that the new network architecture for 5G will be based on software-defined networking (SDN), network functions virtualization (NFV) and cloud technologies, due to their simplicity, flexibility, and openness [5, 8, 9, 10]. SDN is a new network paradigm which centralizes the control and management functions of the network through the decoupling of the control plane and data plane. In this way, the underlying infrastructure just follows the instructions from the controller in the control plane, and data traffic is processed in the granularity of flow instead of packets. Therefore, SDN provides high flexibility and convenience in network management, configuration, maintenance as well as expansion.

In this thesis, we attempt to improve the management efficiency of 5G HetNets by introducing SDN into those networks so as to improve their intelligence and programmability. In SDN-enabled 5G HetNets, the control logic is removed from the underlying infrastructures to a controller in the control layer [8], so that various network functions can be deployed at the logically centralized controller to provide consistent and efficient management over the whole 5G HetNet instead of using particular vendor-specific modules at each BS to perform network functions. With this paradigm, traffic management, including data offloading and load balancing, is optimized and traffic offloading through virtual cluster design in 5G-vehicular ad

hoc networks is also realized in improving the HetNets throughput rate and reducing signaling overhead. Additionally, efficient security provisioning is guaranteed with overall network topology and information sharing under the coordination of SDN.

1.3 Thesis Motivations

5G networks tend to be heterogeneous with multi-tier layout, which provides both high data rate and extensive coverage. However, there are still some critical challenges to be addressed to improve the performance of resource management and security provisioning in 5G HetNets.

HetNets management: Due to the inevitable network densification in the quest for high data rate and the potential mixed use of different generation wireless technologies, 5G tends to have a heterogeneous network (HetNet) architecture with intractable interconnection and limited information sharing among various wireless infrastructures. With an increasing number of infrastructures deployed in future 5G networks, the vendor-specific hardware and protocols would make it challenging and remarkably expensive for operators to adapt their network parameters dynamically. The evolution of 5G networks is thus impeded by the heterogeneity and ossified cellular network architecture with the use of vendor-specific equipment and the limited coordination among co-existing diverse access networks.

Hence, flexible and programmable network management is needed in wireless HetNets to guarantee consistent policy and global management across diverse access networks. In this thesis, software-defined networking (SDN) is introduced into wireless HetNets as an enabling technology to provide high-level control and management. Through the decoupling of the control plane and data plane, SDN enables centralized software control of network functions and policies, such as efficient routing, resource allocation, spectrum sharing, and security provisioning, which also leads to flexible network expansion via BS deployment and easy creation of new applications/ network services. However, despite all these benefits, the performance of SDN networks are still to be modeled and evaluated to validate the advantages, such as SDN processing latency or controller scalability.

Data traffic explosion: The rapidly increased mobile data traffic poses a heavy burden on cellular networks and causes spectrum shortage in licensed bands. Therefore, in this thesis, mobile network congestion is to be alleviated by data offloading using unlicensed band and by optimally balancing loads across multiple cells, while considering network conditions and the quality of service (QoS) requirements of end-user applications.

Additionally, with the anticipated arrival of self-driving vehicles in the foreseeable future, supporting vehicle-related data traffic generated by the expected extensive use of in-vehicle mobile Internet access will become extremely challenging in 5G-vehicular networks. This is mainly due to the high mobility of on-road vehicles, the irregular distribution of vehicles and the complexity of 5G heterogeneous networks (HetNet). How to relieve the traffic burden on 5G-vehicle networks with high mobility is also one of the main topics in this thesis.

Spectrum sharing: Due to the increasing spectrum scarcity caused by data-consuming multimedia applications over mobile and smart devices, efficient utilization of radio spectrum has been receiving tremendous interest during the past few decades. However, existing spectrum sharing techniques mostly use cognitive radio technologies, which rely on the limited sensing capability of devices and lack of timely information exchange between coexisting heterogeneous networks (HetNets). In this thesis, an orchestrated spectrum sharing approach that integrates the distributed located users, base stations (BS), incumbent stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with real-time information exchange will be introduced. To adequately protect incumbent users and efficiently share the pooled spectrum resources, a real-time 3D interference map is to be considered to guide the spectrum access based on the SDN global view.

Security provisioning: Due to the densification of heterogeneous networks and massive deployment of small cells in future 5G networks, users might experience frequent handover and multiple authentications among HetNets due to the reduced cell size. The traditional authentication schemes using three-way handshakes and cryptographic key exchange thus have become inefficient and tend to introduce more latency during frequent handover. Moreover, the simplified access point (AP) makes the storage and calculation of cryptographic keys a challenging

problem. Hence, simplified and efficient authentication schemes that use non-cryptographic methods are to be developed to reduce user authentication frequency and latency, while guaranteeing the secure level of user services at the same time.

1.4 Research Objectives

The research objectives of this thesis are to address the problems mentioned above and challenges in 5G HetNets, namely, data explosion, spectrum scarcity, security provisioning, and HetNets management.

Traffic management improvement: One of the research objectives of this thesis is to relieve cellular network burden through data offloading and load balancing across HetNets but without impacting user application performance. Thus, our particular target is to selectively offload traffic according to different latency tolerance of the user applications to offload cellular network burden using unlicensed bands in Wi-Fi networks, while guaranteeing user QoS at the same time.

Regarding a significant amount of in-vehicle mobile Internet access in 5G-Vehicular Ad Hoc Network (5G-VANET), our target is to introduce SDN as an enabling platform to support the increasing traffic and improve the HetNet management. The moving vehicles are clustered adaptively according to the real-time road topology using SDN's global information gathering and network control capabilities. After that, the vehicles within each cluster communicate with each other using the unlicensed band, e.g., IEEE 802.11p. With the dual cluster head design and beamformed transmission scheme, it is believed that both trunk link communication quality and network robustness will be significantly enhanced.

Spectrum resource utilization efficiency: Given the increasing spectrum scarcity caused by data-consuming multimedia applications over mobile and smart devices, efficient use of radio spectrum is of great importance. Previous spectrum sharing techniques using cognitive radio might cause mis-detection and performance degradation of the incumbent users due to the limited sensing capability of devices and lack of timely information exchange between coexisting

heterogeneous networks (HetNets). Therefore, the objective of this research is to develop an orchestrated spectrum sharing approach that integrates the distributed located users, base stations (BS), incumbent stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with real-time information exchange. To adequately protect incumbent users and efficiently share the pooled spectrum resources, a real-time 3D interference map is to be developed to guide the spectrum access based on the SDN global view.

Simplified authentication and privacy provisioning: Security key management has become difficult in small cells where users join and leave frequently, not to mention the limited capability of simplified access points (APs). Moreover, frequent handover and authentications in small cells also introduce additional latency. Therefore, considering the reduced cell size and simplified energy-constraint small cell infrastructures, simplified and efficient authentication schemes are to be designed to satisfy the more stringent latency requirement in 5G HetNets without harming the secure need of user services.

Our objective of this research is to propose a fast authentication algorithm through the sharing of user-dependent physical layer attributes and a non-cryptographic privacy provisioning method. The SDN platform and fast authentication scheme using weighted secure-context-information (SCI) transfer are also to be designed to improve HetNet management, authentication efficiency during handover, as well as meet the 5G latency requirement.

SDN-enabled HetNets performance: To facilitate the management and information sharing among heterogeneous networks, SDN is introduced into HetNets to add programmability and common interface to the diverse access networks. Through the decoupling of control plane and data plane, SDN enables centralized software control of network functions and policies. That is to say, a script written in the remote data center or the cloud can control the entire network behavior instead of the modules and protocols that have to be implemented at each and every previous BS. To make sure that HetNets performance is improved using SDN technique with no extra latency, SDN-enabled HetNets architecture is to be modeled and evaluated in this thesis using both mathematical analysis and simulations.

1.5 Technical Contributions of the Thesis

The main contributions of this thesis are summarized below:

- To tackle the data traffic explosion challenges, a partial mobile data offloading and centralized load balancing mechanisms are proposed in Chapter 3. To the best of our knowledge, this is the first time that real-time decisions are made for selectively offloading cellular data traffic, while taking quality of service (QoS) of user applications into consideration. Additionally, a two-layered load balancing scheme is proposed to balance the load across cells in network level. The proposed mechanisms are subject to system-level simulations which show an improvement in load balancing, in terms of equilibrium extent and network stability. It is also proved that with the proposed Wi-Fi partial data offloading algorithm, quality of service can be satisfied while saving a significant amount of cellular resources through smart resource allocation.

- A new adaptive clustering scheme with dual cluster head design and beamformed transmission are proposed in Chapter 4 to support the growing in-vehicle data traffic in 5G-VANET. The moving vehicles are clustered adaptively according to the real-time road topology using SDN's global information gathering and network control capabilities. With proposed dual cluster head design and dynamic beamforming coverage, both trunk link communication quality and network robustness are significantly enhanced. Furthermore, adaptive transmission scheme with selective modulation and power control is proposed to improve the capacity of the trunk link between the cluster head and base station. With cooperative communication between the mobile gateway candidates, the latency of traffic aggregation and distribution is also reduced. Matlab simulation results show that the proposed design substantially improves 5G users' bit error rate and trunk link throughput rate.

- To realize efficient spectrum resource utilization, an orchestrated spectrum sharing architecture and a 3D interference map based spectrum sharing algorithm are proposed in Chapter 5. Existing spectrum sharing techniques might cause mis-detection due to the limited sensing capability of devices and lack of timely information exchange between

coexisting HetNets. Therefore, an orchestrated spectrum sharing approach that integrates the distributed located users, base stations (BS), incumbent stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with real-time information exchange is proposed in this thesis. In order to effectively protect incumbent users and efficiently share the pooled spectrum resources, real-time 3D interference map is considered for the first time to guide the spectrum access based on the SDN global view.

- A fast authentication algorithm through the sharing of user-dependent security context information and a non-cryptographic privacy provisioning method are proposed in Chapter 6 to reduce latency and complexity. Existing security key management could be difficult in small cells where users join and leave frequently, not to mention the limited capability of simplified access points (APs). On the other hand, frequent handover and authentications in small cells also introduce unnecessary latency. Therefore, a new SDN-enabled fast authentication scheme using weighted secure-context-information (SCI) transfer is proposed in order to improve authentication efficiency during handover and meet 5G latency requirement. The proposed algorithm is then applied in Neyman-Pearson (NP) hypothesis test to authenticate users, which shows enhanced authentication accuracy and reduced latency in MATLAB simulations. Furthermore, we first analyze the SDN structure using priority queuing theory and prove the performance improvement of SDN-enabled authentication handover.

1.6 Thesis Outline

The rest of the thesis is organized as follows:

In Chapter 2, the research background is briefly given, including the challenges of future 5G traffic management and security provisioning, the core principles of software-defined networking (SDN), the motivation for introducing SDN into wireless networks, the benefits of SDN-enabled HetNets. The chapter also provides a literature survey of SDN and 5G HetNets related to our research.

In Chapter 3, an SDN-enabled application QoS based partial data offloading and load bal-

ancing scheme is proposed for 5G HetNets. In the system model section, the SDN control plane design and SDN-based wireless HetNet model are presented. Given the network topology and consistent policy provided by SDN's global view, the data offloading module is designed, and the partial data offloading algorithm is proposed in terms of application delay threshold. SDN-based load balancing mechanism is also introduced to further balance traffic across neighbor cells. Finally, system-level simulation is conducted in MATLAB to evaluate the performance of SDN structure, novel partial data offloading and load balancing mechanism, as compared to the baseline method.

In Chapter 4, traffic offloading in 5G vehicular ad hoc network (VANET) is discussed, and the adaptive vehicle clustering scheme with beamformed transmission scheme design is proposed. The overall network architecture of SDN-enabled 5G-VANET is given first, and then adaptive clustering scheme is introduced along with dual cluster head design to improve cluster communication robustness. Additionally, after the cluster head is selected, the beamformed adaptive transmission scheme is also discussed in order to guarantee trunk link throughput rate and communication quality. It is also proved by simulation that the proposed schemes provide better throughput rate with lower blocking probability and bit error rate.

An orchestrated spectrum sharing architecture that integrates the distributed systems into an amalgamated network with real-time information exchange and a 3D interference map based spectrum sharing algorithm are proposed in Chapter 5. In order to cope with spectrum scarcity challenges, the limited sensing capability of devices and lack of timely information exchange between coexisting heterogeneous networks (HetNets), an orchestrated spectrum sharing approach that integrates the distributed located users, base stations (BS), incumbent stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with real-time information exchange is proposed in this Chapter. In order to effectively protect incumbent users and efficiently share the pooled spectrum resources, real-time 3D interference map is considered to guide the spectrum access based on the SDN global view.

A new non-cryptographic fast authentication algorithm using physical layer attributes is proposed in Chapter 6. Firstly, the system model of SDN-enabled secure context information (SCI) transfer between 5G equipment is given. Afterward, the weighted SCI design and the user SCI based fast authentication algorithm is explained in detail. The SDN-enabled 5G

privacy protection using partial data transmission is also illustrated. Simulation results witness the improvement in authentication latency as well as the secure level.

Lastly, all the contributions presented in the previous chapters are concluded in Chapter 7. The plan for the future research is discussed in this Chapter as well.

Chapter 2

Enabling Technologies and Challenges of 5G HetNets

2.1 Background of 5G HetNets

2.1.1 Fifth generation (5G) Communication

5th generation mobile networks (5G), also referred to as beyond 2020 mobile communications systems, represent the next major phase of the mobile telecom industry, going beyond the current Long Term Evolution (LTE) and IMT-advanced systems [11]. The ongoing 3GPP RAN meetings have worked to create and to organize the standards of 5G, for example, Polar Code was just decided as the 5G control channel enhanced Mobile Broadband (eMBB) coding scheme and low-density parity-check (LDPC) code was set as the data channel short code in November 2016. In addition to increased peak bit rates, better coverage, and higher spectrum spectral efficiency, 5G systems are required to enable ultra-reliable and low-latency communications (URLLC) and to support Internet of Things with potential numbers of diverse connectible devices, including massive machine type communications (mMTC) devices. In short, 5G should be cost-efficient, flexibly deployable, elastic, and above all, programmable. In order to realize the flexible and elastic deployments and cope with the ever growing mobile data traffic, the mobile infrastructure costs need to be lowered and should be adaptable or expendable. However, the traditional usage of vendor-specific networking equipment can neither

dynamically scale with mobile traffic nor be easily upgraded with new functions.

Along with recent and ongoing advances in cloud computing and their support of virtualized services, it has become promising to design flexible, scalable, and adaptable 5G systems against network ossification by exploiting the virtualization techniques. The advanced main virtualization techniques include network function virtualization (NFV) and software-defined networking (SDN). NFV intends to operate the system on demand on top of fragmented physical infrastructures provided by the networked cloud, while SDN has been seen as a programmable platform to enable scalable and configurable 5G networks. To sum, the high-level 5G network management should be flexible, on-demand and scale up to sudden traffic growth. Meanwhile, the local processing capability should be kept, in the format of cloudlet or the so-called federated networked cloud (i.e., edge or fog computing), to reduce the communication path between end users and servers, as well as the burden of the control plane.

The 5G network system should also allow mobile operators to create and to orchestrate, in a natural and dynamic way, network services targeting a specific optimization of network resources, e.g., traffic offload, security provisioning or optimization of a specific application [11]. Therefore, in this thesis, the SDN-enabled traffic management, and security provisioning are studied, wherein traffic offloading, load balancing, and non-cryptographic authentication applications are specified and enabled exploiting the benefits of common interface across heterogeneous networks provided by SDN. In the following discussion, the background of related technologies and network architectures are investigated and given in detail.

2.1.2 Wi-Fi Technology

With the dramatically increased data traffic demands over the last few decades and in the foreseeable future, heterogeneity of 5G networks has become a necessity with densified deployment of BSs and APs. Therefore, with the flexibility of deployment and the favorable nature of working in the free unlicensed band, Wi-Fi technology has been popular both in 4G communication and is expected to play an essential role in 5G networks to serve more users with increased data rate and flexible coverage at the hotspot.

Wi-Fi is a technology for wireless local area network with devices based on the IEEE

802.11 standards [12]. The IEEE 802.11 standard is a set of media access control (MAC) and physical layer specifications for the implementation of wireless local area network working in the frequency bands of 2.4, 3.6, 5 and 60GHz. From the day it was released, Wi-Fi has been widely used in private homes, businesses, as well as in public spaces for free and convenient Internet access.

Wi-Fi access point (or hotspot) has a range of about 20 meters (66 feet) indoors and a greater range outdoors. Hotspot coverage can be as small as a single room with walls that block radio waves, or as large as many square kilometers achieved by using multiple overlapping access points. Devices that can use Wi-Fi technology include personal computers, video-game consoles, smartphones, digital cameras, tablet computers, digital audio players and modern printers. According to Wi-Fi Alliance, over hundred million people are using Wi-Fi worldwide, and about one thousand million new Wi-Fi devices are used every year [12].

With the popularity of tablets and smartphones such as Apple's iPhone and Google's Android over the last decades, mobile operators have gradually realized that the wireless communication has emerged into an era of mobile data from the era of voice and text. It is commonly accepted that unlicensed spectrum and Wi-Fi can be used as an extension of the fixed broadband networks and deliver the most data services and the best mobile experience at the highest profit margin. Under this market background, Technique companies have been looking for solutions for adding unlicensed band as a complementary in data service provisioning. Qualcomm was the first to propose LTE in unlicensed spectrum (LTE-U), for the use of the 4G LTE radio communications technology in unlicensed frequencies, such as the 5 GHz band used by dual-band Wi-Fi equipment [1]. LTE-U allows operators to boost coverage by using the unlicensed 5 GHz band already populated by Wi-Fi devices. The LTE network keeps a control channel with the mobile device, but all the data flow over the unlicensed band. Ericsson also proposed license assisted access (LAA) technique, which uses a contention protocol known as listen-before-talk (LBT) to regulate the coexistence of LTE with other Wi-Fi devices on the same band. As such, it is worthwhile to investigate the development of Wi-Fi technology and look into the possibilities of Wi-Fi data offloading.

2.1.2.1 Wi-Fi Standards

Among IEEE 802.11 standards, 802.11b, 802.11n and 802.11g use the 2.4 GHz ISM band, operating in the United States under Part 15 Rules and Regulations [1]. Due to the fact that 2.4GHz frequency band is shared by different kinds of users, such as cordless telephones and baby monitors in the United States/ Canada, microwave ovens, and Bluetooth devices, 802.11b and 802.11g equipment may occasionally suffer interference from other users, causing a significant decrease in transmission speed or even service blocking of Wi-Fi signal. That is to say, the quality of service is not guaranteed in Wi-Fi networks, and its service is the best effort based.

During the signal transmission, a Wi-Fi signal occupies five sub-channels in the 2.4 GHz frequency band. Any two sub-channels that differ by five or more in the numbers, such as 3 and 8, do not overlap with each other. For example, channels 1, 6, and 11 are the only non-overlapping channels, as shown in Fig.2.1. Channels 1, 6, and 11 are the only group of three non-overlapping channels in North America and the United Kingdom. In Europe and Japan using Channels 1, 5, 9, and 13 for 802.11g and 802.11n is recommended [1].

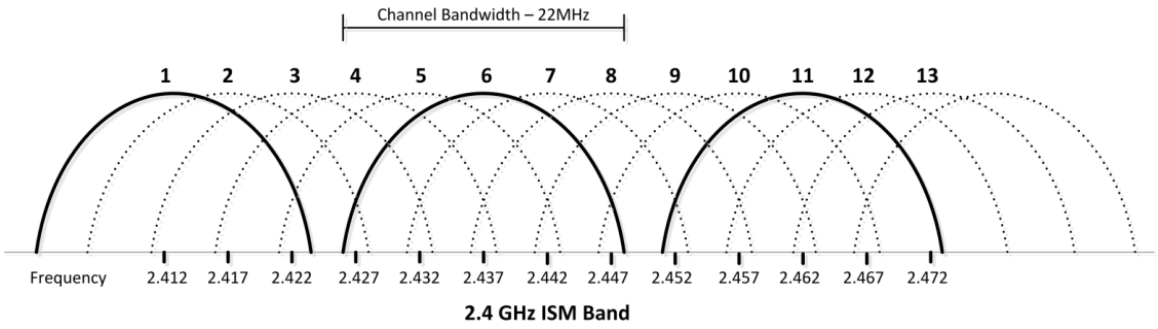

Figure 2.1: IEEE 802.11b/g channels; 1, 6, and 11 are three orthogonal channels [1].

IEEE 802.11a uses the 5 GHz frequency band, which, for much of the world, offers at least 23 non-overlapping channels rather than the 2.4 GHz ISM frequency band [1]. As such, channel hopping becomes a practical and economical solution for mobile devices that are experiencing interference.

2.1.2.2 Wi-Fi Evolution

The first generation Wi-Fi technique has some usability problems. For example, the login process is not convenient. The browser for entering credentials must be done before any service could be used even though the connection manager of the phone indicates connection; secondly, the credentials could be time-limited, and the selection/access of the best hotspots is manual and time-consuming. With its open nature and no physical wire connections, Wi-Fi is also more vulnerable to attacks. As such, next generation hotspot technologies are under development with the use of IEEE 802.11u standard, easy authentication mechanism, and autonomous network selection without human intervention.

IEEE 802.11u has an amendment published by IEEE in 2011, which provides an efficient interface between IEEE 802.11 access network and cellular networks. It has been incorporated into the next generation Hotspot 2.0 specification by Wi-Fi Alliance. To assist user devices to better discover Wi-Fi APs, IEEE 802.11u allows an AP to advertise extra information about its network to the UE via beacon frames. It also allows a UE to query an AP using access network query protocol to discover a range of information, such as roaming partners, IP address type, and so on.

With the new amendment and protocols, Wi-Fi 2.0 (or Hotspot 2.0) can assist users to realize seamless roaming between cellular networks and all the partner Wi-Fi networks [12]. Seamless means that the discovery, selection, authentication and accessing procedure are all autonomous and unnoticeable to users. For example, 3GPP has introduced Access Network Discovery and Selection Function to control the discovery and access of Wi-Fi APs. The function allows a server located in the mobile core network to provide a policy to UE and enable it to only associate with the Wi-Fi AP that meets a predefined threshold. In this way, operators can have control over the selection and association of Wi-Fi APs for specific users.

In this thesis, the control over the Wi-Fi APs is not limited to partner networks. Software-defined networking capabilities are leveraged regarding HetNets information exchange and application level global management to better offload cellular burden while maintaining user QoS.

2.1.3 Vehicle Ad Hoc Networks

It is expected that 5G will be standardized and deployed in 2018 and 2020, respectively. A key scenario for 5G is connected mobility, and one of the most important parts of connected mobility is vehicular communication for infotainment, safety, and efficiency [13]. However, vehicular communication for the above purpose is sensitive to latency and signal reliability. In the following discussion, vehicular ad hoc network (VANET) will be introduced, and the challenges in vehicle induced data traffic transmission are summarized.

The Vehicular Ad hoc Network (VANET), is a technology that is created from the principle of mobile ad hoc network (MANET) and uses moving cars as nodes in a network to create a mobile network. The participating vehicles are equipped with wireless transceivers and can exchange data with neighbor vehicles around 100 to 300 meters of each other or road side units (RSUs). In this way, the participating vehicles work as wireless routers and create a network with a wide range. Such architecture provides the potential convenience for improving the road safety, congestion prediction and avoidance through information exchange, accident monitoring, as well as in-car information and entertainment systems.

As shown in the Figure 2.2 above, VANET architecture mostly includes vehicles moving on the road that are embedded onboard units, RSUs and a server (e.g., traffic management server) that are in charge of the resource allocation of the distribute RSUs. RSUs serve as the gateway between vehicle networks and other networks or agents, such as the Internet and the remote control server. It can also be seen from the figure that there are two types of links in VANET: vehicle-to-vehicle (V2V) communication and vehicle-to-infrastructure (V2I) communication. For V2V communication, vehicles exchange message with neighbor nodes in an ad hoc manner without a high-level coordination. In V2I communication, vehicles share the message with RSUs.

VANET makes high-speed Internet access available in the car and gives traveling more “live times”. In the earlier stage, such a network is doubted as it poses safety concerns for careless or distracting driving. With the advancement of artificial intelligence and sensor technologies, autonomous vehicles can sense their surroundings in real-time by the combined use of many techniques including radar, lidar, GPS, Odometer, and computer vision. Consequently,

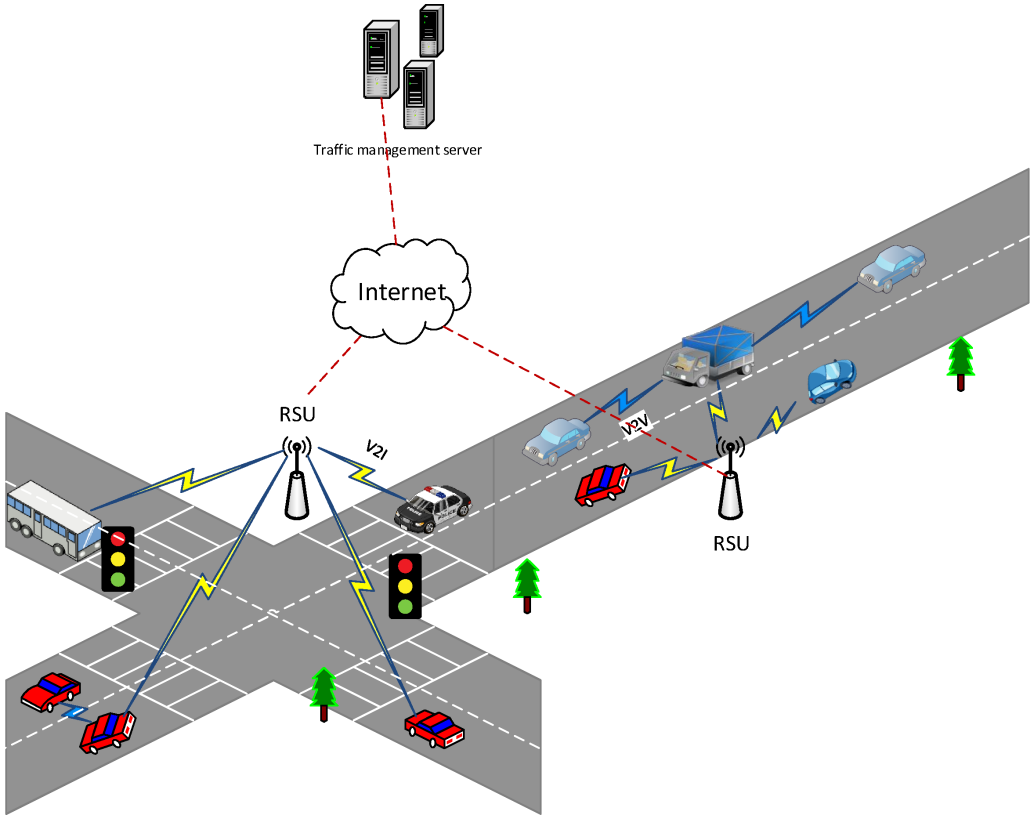

Figure 2.2: Vehicle ad hoc network scenario.

the self-driving vehicle is closer to reality than we ever thought. According to recent market survey [14], it is envisioned that more than 15 million self-driving cars will be on the road by 2030. We can predict that very soon drivers would be released from the burden of driving and thus have time for mobile Internet access, such as VoIP with family, watch news highlights or entertainment.

Therefore, academia and industry become interested in the development of VANET technology, especially since 1990*s*. Their efforts have led to the introduction of a physical layer standard based on IEEE 802.11 operating in the 5.9 GHz band in the ad hoc mode. The physical layer standard, namely, Wireless Access in Vehicular Environments (WAVE) / Dedicated Short Range Communications (DSRC) standard, has been formalized and published by the IEEE as 802.11p standard for vehicle communication [15]. At the same time, security and application standard is also under development for VANETs on top of IEEE 802.11p by IEEE

1609 working group [16].

However, VANET has to operate in a challenging environment due to high mobility nature of vehicles and highly variable channel conditions. IEEE 802.11p protocol was first proposed for V2V communication without the assistance of RSUs or BSs. The carrier sense multiple access (CSMA) mechanisms might cause collision and performance degradation, especially during rush times. The practical implementation difficulty is further compounded by the ad hoc nature of VANET due to the lack of optimal channel access and quality guarantee. Borrow from the idea of MANET clustering algorithms, VANET clustering algorithm works by grouping fast moving vehicles (which is different from MANET) into clusters according to some predefined rules and selecting one of them as the cluster head (CH) to mediate between the cluster and the other part of the network. The maintenance of the VANET clusters requires channel monitoring, mobility prediction through machine learning, and security consideration. Authors in [17] proposed a multihop clustering algorithm through monitoring the neighbor topology and thus extend cluster lifetime. In a recently published survey paper [18], the authors explore the design considerations of clustering algorithms in VANET, especially in the area of cluster head selection, cluster affiliation, cluster management, and identifies new directions and recent trends in the design of these algorithms. The paper also reviewed the methodologies for validating the clustering performance, and point out the main shortcoming is the lack of realistic vehicular channel modeling.

In summary, most of the cell head selection and cluster affiliation algorithm by far rely on the interaction between nearby vehicles. For example, upon entering a VANET, the vehicle will announce its existence to its neighbors using a periodic broadcast and gather the location information from its neighbors. Afterward, all the vehicles will access its suitability as a cluster head (CH), and the other vehicles will decide the affiliation to this cluster. However, it is believed that high-level coordination is critical in the vehicle clustering procedure as the QoS of user traffic and the spectrum efficiency is not able to be guaranteed in ad hoc way. Moreover, in next generation 5G communications, the information exchange among heterogeneous network infrastructures is essential for the prediction of road traffic topology due to the reduced cell size.

Therefore in this thesis, the coordination and information exchange between 5G HetNets

are investigated, and the vehicle clustering scheme, beamformed trunk link transmission are also analyzed in detail.

2.2 5G HetNet Management Platform-Software Defined Networking (SDN)

2.2.1 Fundamental Idea of SDN

The traditional Internet devices are usually vendor-specific with a control plane, data plane, and a management plane. Control plane realizes the control functionality of packet forwarding, such as updating forwarding tables through layer two learning in switches, or updating the routers' routing tables through IP protocols. Additionally, the control plane produces access control list, either dynamically or statically, according to user configuration. Data plane thus processes packets, look-up tables and maps the incoming packets to outgoing ports, in accordance with the instructions from the control plane. The management plane manages the network configuration through command line interface (CLI) and Simple Network Management Protocol (SNMP).

The Internet has been designed in a way that transports only rely on core network functions despite the application needs, such as prioritize sub-flows that carry specific objects. Internet service is summarized as best-effort service. However, this does not mean that information about the transport/application requirements would not be helpful to improve the efficiency of the network or to enable the application to receive the most suitable service [19]. With the ever-increasing traffic demands and expanding network sizes in recent years, the complexity of protocol implementation on diverse network devices and the difficulty to add/ remove any devices become a burden for Internet innovation and hinder network development [20]. A new transport protocol has to traverse the network to become usable. However, the ubiquity of middleboxes of a variety of forms (from firewalls, accelerators, load balancers, and a range of portals and more exotic devices) makes it very hard to change the status quo after set up [19]. It is hard to configure a device without touching multiple switches, routers, firewalls, web authentication portals, and other protocol-based mechanisms through various device-specific

management tools. In addition to the configuration problem, challenges also exist in link failure recovery latency, namespace management, and hard to debug without log information.

The idea of Software-defined networking (SDN) began shortly after the release of Java in 1995, and has received great attention as an enabling solution to simplify Internet management since 1998 [21]. The principal concept behind SDN is to separate the control logic from network equipment to control layer. In this way, the underlying infrastructures are simplified, and they only follow instructions from the controller through the southbound interface (e.g., OpenFlow [22], etc.). The advantage of this layered network structure is apparent. Firstly, a network topology can be gathered at the controller for policy making such as routing decisions, network traffic monitor, and pre-judgment. Secondly, network protocols and other applications could be written to the controller with real-time interaction. Thirdly, as devices are simplified to just follow instructions, network innovation is much easier. New functions can be added in software instead of modifying each switch. As a result, enterprises and carrier operators gain unprecedented programmability, automation, and flexible network control, enabling them to build highly scalable, flexible networks that readily adapt to changing business needs.

2.2.2 SDN Evolution

Strictly speaking, SDN is not a revolutionary new networking technology. The idea of programmable network and separated control data plane can be traced back to ancient programmable networks, for example, Open Signaling [23], Active Networking [24], and Devolved Control of ATM Networks (DCAN) [25].

Open Signaling was proposed in 1995 as a technology that aims at making ATM, the Internet and mobile network more open, extensible, and programmable [23]. The motivation of Open Signaling is that decoupling of control software and hardware is complicated due to closed nature of network infrastructures. Therefore, programmable network interfaces need to be set up to provide access to different hardware to make the network as programmable as the personal computers.

Active Networking was introduced in the 1990s, which allows the possibility of highly tailored and rapid real-time changes to the underlying network operation. Active Networking

allows data to change its form (code) to match the channel characteristics as the code is sent along with information packets. It also enables the use of real-time customized programs within the network to compose network services. However, the code compression and computation remain a challenge for active networking.

DCAN was also introduced in the 1990s. As seen from the name, DCAN provides an infrastructure for scalable control and management for ATM networks [25]. The principle of DCAN is similar with that of SDN: control functions in ATM switches should be separated from devices and gathered to an external controller.

Due to the mismatch with Internet development of that age and their inherent limitations, most of these technologies were not widely accepted or used by industry. With the development of cloud computing and the expanding of Internet, the network ossification becomes more and more unbearable, and the quest for innovation brings programmable network concept to the front of academy and industry again. Since 2004, 4D Project [26], NETCONF [13], and SANE/Ethane [15] project were proposed consecutively and have finally led to the technology that is based on advanced computing and virtualization capability: Software-defined Networking.

2.2.3 SDN Architecture

Both Open network Foundation (ONF) and Software-Defined Networking Research Group has investigated SDN from various perspectives. Open network foundation is a non-profit organization founded by Google, Microsoft, Yahoo and some Telecom Operators in March 2011, which aims at the development of SDN-related technology, standardization, and marketing. In reference [20], the structure and definition of SDN are provided as follows:

"In the SDN architecture, the control and data planes are decoupled, network intelligence and state are logically centralized, and the underlying network infrastructure is abstracted from the applications."

Fig. 2.3 above also shows the architecture of software-defined networking. It is clear that SDN is a structure that separates the control plane (or network operating system) from the forwarding plane, which was previously vertically integrated into traditional devices. In SDN, the

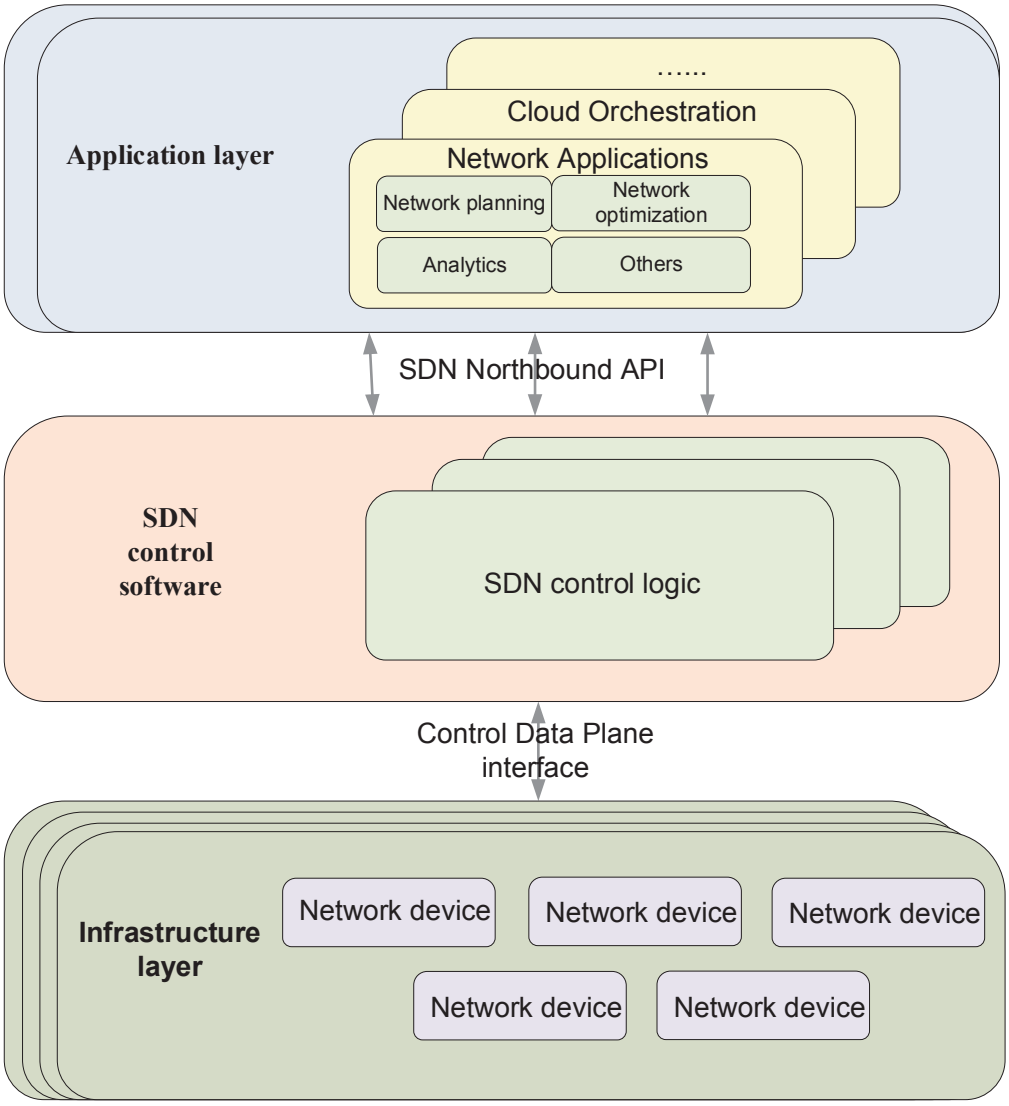

Figure 2.3: The architecture of Software-defined Networking.

controller observes and gathers the entire network state from a central vantage point, hosting features such as routing protocols, access control, network virtualization, energy management, and so on. As the controller is a program running on a server, it can be located anywhere and communicate with network devices through southbound protocols such as OpenFlow protocol through secured channel. With the centralized control, switch state information can be collected to form an overall network topology and thus ensure real-time management and consistent policy. There has been continuous work on the development of SDN controllers, and the leading examples are NOX, ONIX, Beacon, Maestro, Floodlight [21], etc. Although the

idea of separation has been around in previous technologies mentioned before, the nascent part of SDN is the complete separation, OpenFlow control protocol, and the design space it can provide for network operators.

2.2.4 Control-date Plane Interface

Control-date plane interface, namely, southbound interface and protocols, facilitates efficient control over the network devices and enables the SDN Controller to make changes according to real-time demands and updates dynamically. OpenFlow, which was developed by the Open Networking Foundation, is the first and probably most well-known southbound interface protocol. It is an industry standard that defines the way the SDN Controller interacts with the forwarding plane and make adjustments to the network, to better adapt to the changing business requirements [27]. With OpenFlow protocol, entries can be added and removed to the internal flow-table of switches and potentially routers to make the network more responsive to real-time traffic demands.

To be specific, OpenFlow could be implemented on both side of the interface (i.e., devices interface and controller interface) by simple firmware or software upgrade. It defines how traffic goes based on usage patterns, applications requirements and resource allocation in per-flow basis, which makes it adaptable to real-time traffic variation. OpenFlow defines the forwarding rule for each flow and performs the corresponding actions accordingly (e.g. drop, forward, modify, or enqueue). Moreover, multi-level flow or flow line has been developed to improve routing efficiency.

In the idle state, the controller just sends Echo message itself. When the first packet of an arriving data flow comes, and there is no flow entry in the switch, the first packet will be sent to the controller for routing decision. The controller then checks the security policy to make sure the request is legitimate. If so, the controller looks up the flow table to see if there is history to refer to or broadcasts the request and computes a path for the arriving flow. Afterward, the path would be pushed back into switches along the route. Next time when same type data flows arrive, the controller will push the path directly until the stored flow table time-out.

As shown in Fig. 2.4 above, flow table includes match fields, rules, and action. The rule

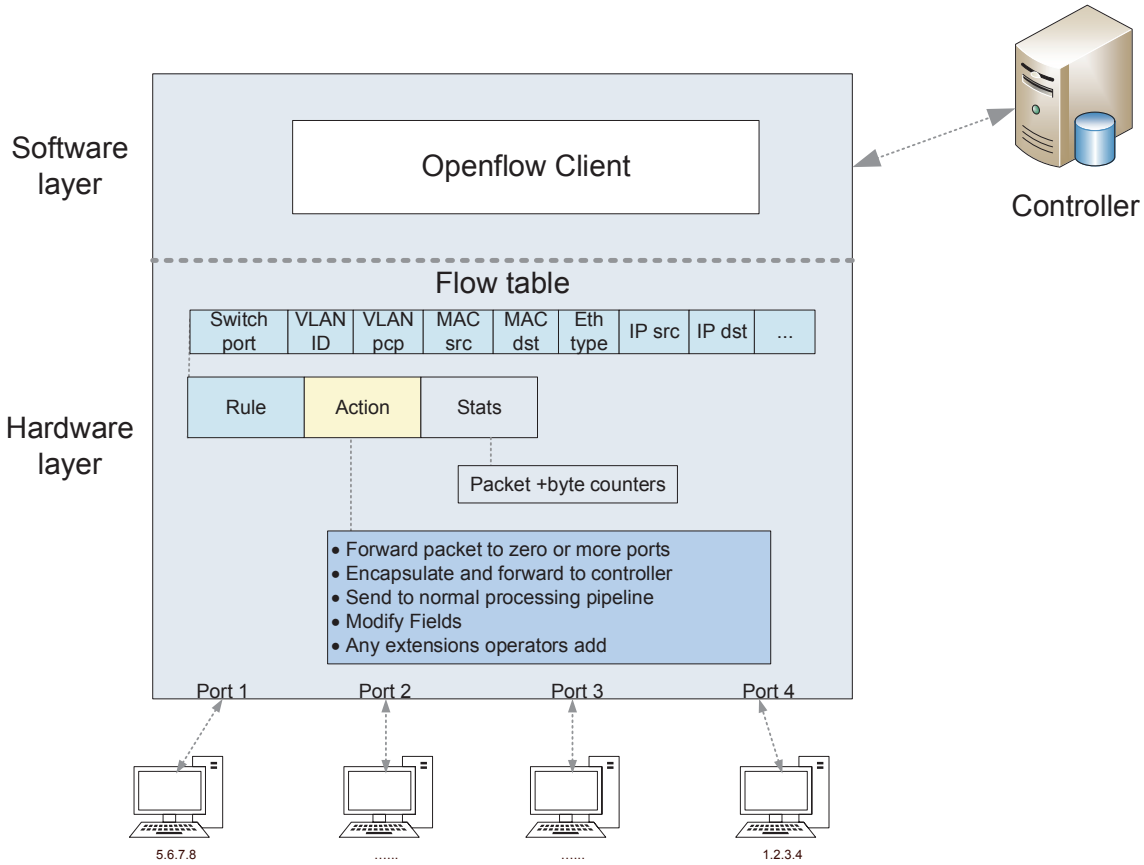

Figure 2.4: The flow table structure of OpenFlow protocol.

defines the flow, which consists of the mainly field in the packet header. Each header field can be a wild-card to allow for aggregation of flows. Identification of link layer, network layer, and transport layer are in match fields. The action defines how the packets should be processed, while Statistics keep track of the number of packets and bytes for each flow, and the time since the last packet matched the flow (to help with the removal of inactive flows). Each flow entry in the flow table of the switch has a simple action associated with it. For example, forwarding, encapsulating and dropping. Forwarding a flow's packets to a given port allows packets to be routed through the network according to pre-defined rules. Encapsulate and forward a packets to the controller mean that a packets, typically the first one in a new flow, is encapsulated and forwarded to the controller through secure channel, so that controller can decide if the flow should be added to the Flow Table. Packets are dropped when there is security consideration, i.e., to curb denial of service attacks, or to reduce spurious broadcast discovery traffic from end-hosts.

A number of switch and router vendors have announced their support of OpenFlow, including Cisco, Juniper, Big Switch Networks, Brocade, Arista, Extreme Networks, IBM, Dell, NoviFlow, HP, NEC, among others. Although OpenFlow is the most well-known of the SDN protocols for the southbound interface, it is not the only one available or in development. The Network Configuration Protocol (NetConf) uses an Extensible Markup Language (XML) to communicate with the switches and routers to install and make configuration changes; Lisp, also promoted by ONF, is available to support flow mapping. In addition, there are more established networking protocols finding ways to run in an SDN environment, such as Open Shortest Path First (OSPF), Multi-protocol Label Switching (MPLS), Border Gateway Protocol (BGP), and so on.

2.2.5 SDN in Wireless Network

Existing cellular networks suffer from vendor-specific equipment/configuration interfaces, inflexible and expensive infrastructures, and complex control-plane protocols. Therefore, related studies have been working on the simplification of the design as well as the management of cellular networks since the earlier work in 2010 [28]. In [29], SDN architecture with extensions to controller platforms, switches and base stations are proposed for mobile networks to simplify the network design and improve the real-time, fine-grained management. In [30], authors proposed SoftCell, which supports fine-grained policies for cellular core networks. Based on SoftCell, packet classification on base station's access switches and the aggregation of traffic along multiple dimensions were also designed to improve the scalability and flexibility of core network. K. Pentikousis et al. in [31] presented the SDN-based mobile carrier network architecture which can provide flow-based forwarding model and a platform for innovation for operators instead of end up in vendor lock-ins. The performance is also validated by testbed implementation and experimentation.

Considering the network overhead and quality of service, authors in [32] aims at minimizing the transport network load and infringed data-plane delay imposed by introducing the new emerging technology-Network Functions Virtualization (NFV) and Software Defined Networking (SDN). A model that resolves the services placement problem was proposed, where

NFV and SDN are applied selectively with the consideration of data-plane delay, the number of potential data centers and SDN control overhead. The proposed concept was also validated through use case examples. In [33], the key reasons for the transition to SDN-based mobile networks and implementation proposals of design scenarios were described. Authors also emphasize SDN's contribution to more efficient inter-cell interference management, traffic control and network virtualization. As one of the most famous SDN-related projects detailing with mobile cellular networks, SoftRan is designed as a rethink of radio access layer in [34]. SoftRan is a software defined centralized control plane for access networks, which abstracts all base stations in the local geographical area as a virtual big base station managed by the centralized controller and perform interference management, load balancing, with the consideration of the overall network view. Authors in [35] also tried to investigate centralized and distributed radio resource management (RRM) architectures for the realization of Enhanced intercell interference coordination (eICIC). The performance comparison was also given, where the centralized RRM structure assumes macrocells and remote radio heads (RRHs) interconnected via high-speed fronthaul connections, and the distributed architecture is based on traditional macrocell and picocell deployments.

However, none of the aforementioned approaches investigate the mathematical model of the SDN controller and give a discussion about the advantage, disadvantage brought by SDN-based model design with a centralized Radio Access Network (RAN) controller to manage eNodeB applications/functionalities of the heterogeneous cellular networks. In the following chapters, the SDN control applications are designed for data offloading, fast authentication, the performance of the centralized SDN structure will also be analyzed to verify the improvement.

2.3 Resource Management and Security Provisioning in 5G HetNets

2.3.1 Resource Management

Heterogeneous networks (HetNets) with a mixed deployment of macro cells and small cells such as picocell, femtocell, and relay node over existing network architecture have been ex-

pected to play a crucial role in the next generation 5G cellular networks. Through densified deployment, HetNets improves the signal quality by reducing the distance between base stations (BS)/ access points (AP) and the user equipment. Additionally, small cells can enhance the network throughput by providing high data rate at busy areas, while macro cells guarantee the wide coverage.

Despite the advantage of more frequency reuse due to smaller cell coverage, small cells also bring network management challenges, especially regarding the coordination between a multitude of wireless systems, traffic management and resource allocation among HetNets. Therefore in the thesis, traffic management schemes in 5G HetNets, especially data offloading and load balancing design will be studied as an essential part to support 5G data traffic requirements. Next, the literature review of traffic management schemes will be given in detail.

Data offloading Data offloading is one of the most widely accepted traffic management schemes in addressing spectrum shortage concerns through the unlicensed band. Offloading refers to the use of complementary networks for delivering the data initially targeted for the cellular networks [36]. As shown in [37], 55% of the overall data usage occurs at home, and 26% of them happens in office or hotspots. Therefore, the already deployed Wi-Fi access points (APs) in indoor scenarios or commercial hotspots become a natural and economy solution for the operators to execute data offloading.

As almost all the smartphones nowadays have built-in Wi-Fi card, Wi-Fi presents an attractive offloading technology for the operators with its widespread use and its ability to shift data traffic from the expensive licensed bands to the free unlicensed bands (e.g., 2.4 GHz and 5 GHz). Therefore, many related works have been done on Wi-Fi based offloading. For example, the algorithm presented in [38] predicts the future Wi-Fi connectivity and delays suitable data transfers until a Wi-Fi network becomes available. Sou *et al*. In [39] propose a more flexible Wi-Fi offloading model, by introducing mobile Internet Protocol (IP) flow into the core network. The authors in [40] and [41] present a distributed offloading solution based on non-cooperative game theory, where macro cellular BSs and third party APs try to achieve the highest utility through traffic offloading.

All of the works above show significant performance improvements regarding traffic man-

agement. However, technical challenges still exist, especially when the envisioned 5G HetNet architecture is taken into consideration. Firstly, the uncoordinated Wi-Fi cells will be deployed overlay to the heterogeneous cellular cells [42], resultantly, resource management will be challenging in this two-tier architecture. Secondly, offloaded data will be routed directly to the Internet through the Wi-Fi backbone, which is not under control of the wireless operators since the Wi-Fi networks ae usually owned by third parties. Consequently, the operators are unwilling to lose control over the valuable subscriber information [3], not to mention the security risks introduced due to the loosely controlled nature of the Wi-Fi networks. Moreover, the previous data offloading schemes lack consideration about user QoS requirement and the different delay constraints of the traffic. Therefore, it comes as no surprise that operators are seeking to introduce intelligence while exerting greater controls in 5G HetNet traffic management [8], and also design better offloading/load balancing schemes based on the controller's global view of the HetNets.

Additionally, as two major 5G dimensions, mMTC and URLLC correspond to the large number of machine type communication, self-driving, and Internet of vehicles. Therefore, traffic management in vehicular ad hoc network (VANET) is also discussed in this thesis, especially regarding in-vehicle traffic offloading and communication quality. VANET is a technology that uses moving cars as nodes to create a mobile network, allowing cars approximately 100 to 300 meters of each other to connect and, in turn, create a network with a wide range. VANET offers benefits through the connection of vehicles and roadside units. Imagine that in the near future, vehicles communicate with each other to avoid accidents and traffic jams; nevertheless, traffic light systems are also included to reduce redundant stops and optimize the traffic signal control system. It is undoubted that the era of connected vehicles will arrive shortly and related studies would be highly essential in 5G networks.

Regarding 5G-VANET in-vehicle traffic management, there have been lots of recent research in vehicle clustering and the introduce of IEEE 802.11p protocol as the intra-cluster communication protocol to reduce cellular network burden. Authors in [43] provide a clustering method which considers vehicle velocities, the direction of movement, and inter-vehicle distance. However, it fails to take the unpredictable nature of vehicle mobility into consideration. In [44], a dynamic clustering-based mobile gateway management mechanism is proposed,

which considers vehicle movement and executes clustering algorithm periodically. However, how to decide this period remains an open problem, and the maintaining of the cluster could dramatically increase the computing burden on cluster head. With a coexistence of multiple HetNet infrastructures in the future 5G network, it is also difficult for a single base station to predict the arriving traffic and execute clustering algorithms adaptively due to the limited interconnection.

Load balancing In addition to spectrum shortage, network congestion is another critical challenge for the cellular networks. Congestion results from the extremely large number of bandwidth-hungry smart devices. To alleviate network congestion, load balancing, which balances data distribution across multiple resources, e.g., multiple macrocells, presents an attractive solution [45]. Thus, offloading, along with load balancing, are used to address spectrum shortage and network congestion issues, and they optimize the resource utilization in cellular networks.

Load balancing [45] can reduce network congestion over an area by distributing user traffic across neighboring APs or BSs. With load balancing, a proportionate share of wireless traffic can be guaranteed for better resource utilization. Since it is hard to guarantee QoS with Wi-Fi, load balancing and vertical handovers of the edge users are seen as the enabling solutions to tackle network congestion. Both these strategies improve QoS by equally distributing the traffic load across the network [36]. In load balancing related research, a basic distributed mobility load balancing algorithm is introduced in [45]. *Haydar et al*. studies access selection between heterogeneous networks by considering QoS and present an algorithm which improves load balancing performance in [46]. However, existing load balancing algorithms are mostly distributed, which causes ping-pong handovers due to the lack of global information [47]. Therefore, better load balancing scheme is also to be designed with consideration of the HetNets environment.

2.3.2 Security Provisioning

More efficient security provisioning mechanisms are also essential for 5G networks due to the increased latency requirement. Challenges also come from the fact that on one hand, densifi-

cation of heterogeneous networks and massive deployment of small cells become the natural choice of 5G and render frequent handover across cells; on the other hand, many applications supported by 5G like mobile banking and cloud-based social applications require even higher data confidentiality and reliable authentication against malicious attacks.

The common practice for secure communications in 3G and later wireless networks is based on admission control and cryptographic exchange. Fig. 2.5 gives an overview of the handover authentication procedures between heterogeneous networks and within the same network [2]. The associated network components here are the user equipment (UE), the base station (BS) or access points (AP), and an authentication server. It can be seen from Fig. 2.5 that mutual authentication during handover between the user and a new network, i.e., procedure 1, is realized by the pairing of particular hashing output. Each time the involved vector includes RAND (a random number known by the server), AUTH (authentication token sent by server), Pairwise key, etc. For mobility within the same network, i.e., procedure 2, the current serving AP will inform the target AP about the possible handover so that the latter can retrieve the user authentication and key context from the server. In the following, we analyze existing handover authentication procedures and identify the challenges in 5G HetNet based on Fig. 2.5.

To enable handover between different wireless networks (i.e., procedure 1 in Fig. 2.5), various authentication servers and protocols are involved due to the closed nature and structure of each network in HetNet, rendering frequent establishments of the trust relationship and authentications during mobility, especially in 5G small cells scenario [9]. The 3GPP committee has provided particular key hierarchy, and handover message flows for various mobility scenarios [48]. However, the specific key designed for handover and different handover procedures for different scenarios will increase handover complexity when applied to 5G HetNets. As the authentication server is often located remotely, the delay due to frequent inquiries between small cell APs and the authentication server for user verification may be up to hundreds of milliseconds [49], which is unacceptable for 5G communications. Authors of [50, 51] have proposed simplified handover authentication schemes, i.e., direct authentication between UE and APs, based on public cryptography. These schemes realize mutual authentication and key agreements with new networks through a 3-way handshake without contacting any third party, like authentication, authorization, and accounting (AAA) server. Although the handover

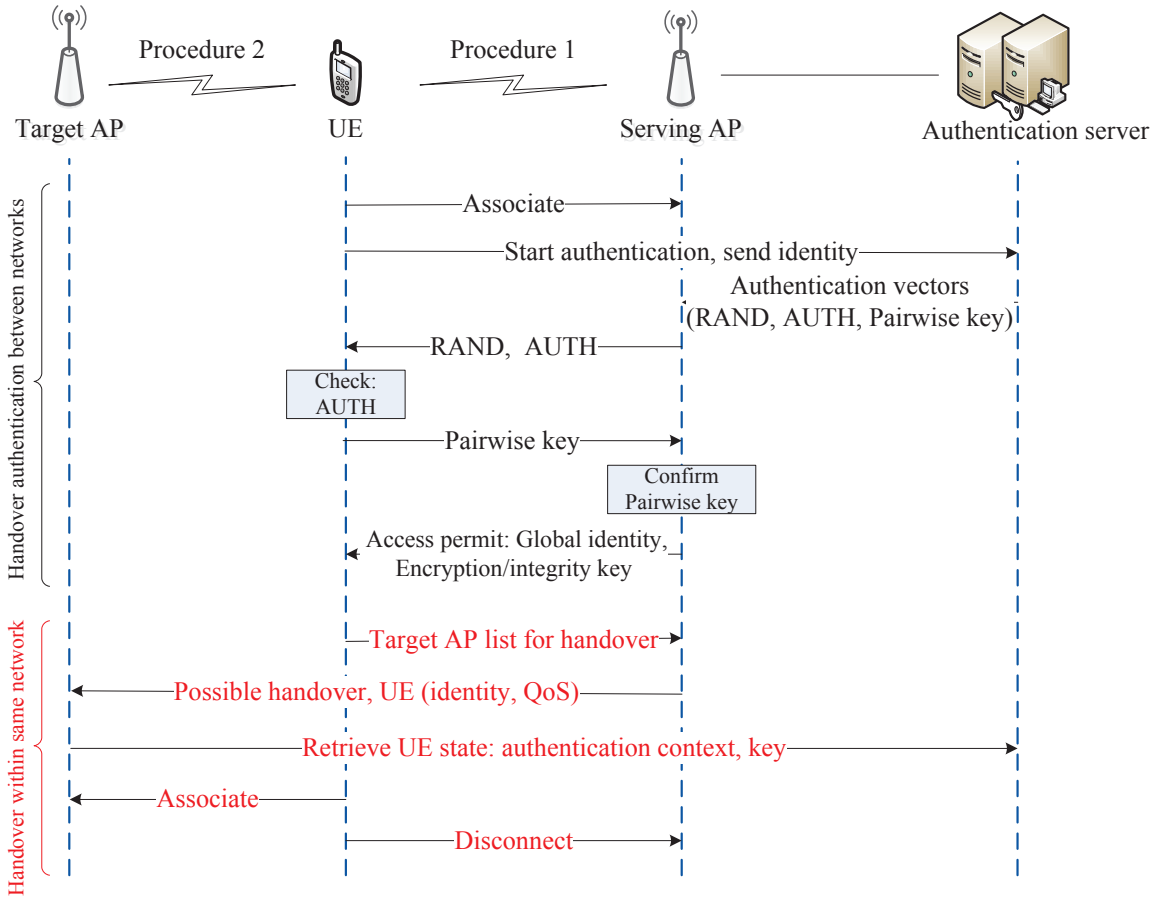

Figure 2.5: Authentication processes of handover procedure 1: between different networks and handover procedure 2: within a same network [2].

authentication procedure is simplified, computation cost and delay are increased due to the overhead for exchanging more cryptographic message through wireless interface [49]. Due to the same reason, carrying a digital signature is secure but not efficient for dynamic 5G wireless communications.

For handover within the same network (i.e., procedure 2 in Fig. 2.5), existing security mechanisms utilize complex context transfer, and it has been found that most of the handover latency is due to the scanning time for identifying the target AP and round trip time to the authentication server. Related work in [52] proposed a user-assisted authentication context transfer scheme, by which the current AP transfers signed authentication certificate as security context to the user, and then to the target AP through the user. The UE is actively involved in handover authentication with its existing connections with the current and next target AP to reduce latency. However, mutual trust between APs is assumed in these solutions, which could

be infeasible for 5G HetNet due to the lack of direct interface between different networks. Also, the transferred security context, which is just a combination of identity and signature, may not be secure enough to prevent 5G wireless communication from potential attacks.

In light of these challenges, robust and efficient handover authentication and privacy protection are crucial in securing 5G networks. For example, the unique link characteristics experienced by each UE can be explored as security context to accelerate authentication handover. Such user specific attributes include physical layer characteristics (clock skew, signal strength, channel state information), location, even moving speed and direction [53], some of which have already been reported to APs for the purpose of resource allocation and seamless handover. It is believed that by taking advantage of these unique attributes combinations as non-cryptographic solutions, authentication can be faster, more robust and less complex compared with widely used cryptographic exchange mechanisms [54].

2.4 Chapter Summary

In this chapter, the background concepts and models used in the thesis are reviewed. At the beginning of this chapter, a brief introduction of 5G communication is given, emphasizing the quest for flexible and expandable network infrastructure and the challenges of existing vendor-specific heterogeneous networks. After that, virtualization techniques are introduced to design flexible, scalable and adaptable 5G systems against network ossification. Based on the virtualization control model, 5G systems can allow operators to create and orchestrate network services dynamically. Therefore, the background and literature review of 5G traffic management and security provisioning are also presented. From Section 2.2, software-defined networking, which is used in the later discussions in the thesis, is discussed. Both the concepts, architecture and related works are given, especially SDN design in mobile networks.

Chapter 3

SDN-enabled Data Offloading and Load Balancing in 5G HetNets

3.1 Introduction

Over the last decade, the cellular networks have witnessed an unprecedented growth in data traffic and the diversity of service in mobile networks, mainly due to the explosive development of smart and media-rich mobile devices. These smart devices enable ubiquitous mobile Internet access, traffic-intensive social applications and cloud-based services [3]. According to Cisco's networking visual index report [6], the data traffic has reached 7.2 exabytes per month at the end of 2016, up from 4.4 exabytes per month at the end of 2015. The mobile data traffic is predicted to grow at a compound annual growth rate of 47 percent from 2016 to 2021, reaching 49.0 exabytes per month by 2021. Current cellular networks are not capable of sustaining such high traffic volumes. Moreover, the updating of the cellular network infrastructures can not be as fast as the growing of mobile data traffic. Therefore, the fifth generation (5G) is being touted as the next-generation cellular standard and operators have been evolving wireless network infrastructures to provide increased data rates and deploying small cell solutions to meet the increased demand, with the aim to enable a fully connected society. The 5G networks are envisioned to have a heterogeneous network (HetNet) architecture, combining multiple radio access technologies (multi-RATs) into a single holistic network [9]. To optimize resource utilization and to support the predicted traffic volumes, 5G is expected to utilize a multitude

of technologies including the device to device (D2D) communications, data offloading, load balancing, spectrum sharing, etc.

Data offloading is seen as an enabling solution to address the spectrum shortage concerns. Offloading refers to the use of complementary networks for delivering the data initially targeted for the cellular networks [36]. As shown in [37], 55% of data usage occurs at home, and 26% occurs in office or hotspots. Thus, the already deployed Wi-Fi access points (APs) become a natural solution for the operators to execute data offloading. The operators have been offloading traffic to Wi-Fi APs and deploying their AP infrastructures. In addition to spectrum shortage, network congestion is another critical challenge for the cellular networks. Congestion results from the extremely large number of bandwidth-hungry smart devices. To alleviate network congestion, load balancing, which balances data distribution across multiple resources, e.g., multiple macrocells, presents an attractive solution [45]. Thus, offloading, along with load balancing, are used to address spectrum shortage and network congestion issues, and they optimize the resource utilization in cellular networks.

With the built-in Wi-Fi card in the smartphones, and the ability to shift data traffic from the high-cost licensed bands to the free unlicensed bands (2.4 GHz and 5 GHz), Wi-Fi presents an attractive offloading technology for the operators. Therefore, many works have been done on Wi-Fi-based offloading. For example, the algorithm shown in [38] predicts the future Wi-Fi connectivity and delays suitable data transfers until a Wi-Fi network becomes available. Sou *et al*. in [39] propose a more flexible Wi-Fi offloading model, by introducing mobile Internet Protocol (IP) flow into the core network. The authors in [40] and [41] present a distributed offloading solution based on non-cooperative game theory, where macro cellular BSs and third party APs try to achieve the highest utility for offloading traffic. Authors in [55] proposed a matching algorithm for the assignment of the offloaded mobile devices to the APs, to minimize the average delay of the offloaded packets as one of the main QoS parameters on the one hand and to maximize the utility of the APs on the other hand. In load balancing related research, a basic distributed mobility load balancing algorithm is introduced in [45]. *Haydar et al*. study access selection between heterogeneous networks by considering QoS and present an algorithm which improves load balancing performance in [46].

All of the works above show significant performance improvements regarding traffic man-

agement. However, technical challenges still exist, especially when the envisioned 5G HetNet architecture is taken into consideration. Firstly, the uncoordinated Wi-Fi cells will be deployed overlay to the heterogeneous cellular cells [42], resultantly, resource management will be challenging in this two-tier architecture. Secondly, offloaded data will be routed directly to the Internet through the Wi-Fi backbone, which is not under control of the wireless operators since the most of the Wi-Fi APs are usually owned by third parties. Deploying Wi-Fi APs by operators themselves is not cost-effective nor energy-efficient. Consequently, the operators are unwilling to lose control over the valuable subscriber information [3], not to mention the security risks introduced due to the loosely controlled nature of the Wi-Fi networks. Moreover, the QoS of the data traffic is not able to be protected in Wi-Fi networks with the contention-based medium access, especially if mobile devices just "offload' to Wi-Fi in preference to cellular. All of these drawbacks make the coordination, integration and timely information exchange of heterogeneous networks critical for data offloading, in particular between mobile network and Wi-Fi network. On the other hand, existing load balancing algorithms are also mostly distributed, which causes ping-pong handovers due to the lack of global information [47]. Therefore, it comes as no surprise that operators are seeking to introduce intelligence while exerting greater controls in 5G HetNet traffic management [8].

This chapter explores Software-Defined Networking (SDN) [20], a programmable network structure, as an enabling solution to apply intelligence and control in 5G HetNets [8]. The idea of programmable networks has been around for many years. Nevertheless, the advent of the OpenFlow interface has given a new life to SDN [21]. OpenFlow was first introduced in [22], where the authors provide a uniform interface for researchers to program flow entries and to run experiments on Ethernet switches, without having any knowledge of the internal workings of the switch. In the context of wireless communications, the authors in [56] introduce OpenRoads-an open SDN platform which improves mobility management in HetNets. Related work in [57] further provides an SDN approach for handover management in heterogeneous networks, and the real-time test bed shows significant performance improvement in the QoS of the real-time videos.

In this chapter, we propose SDN-based partial data offloading and load balancing algorithms to alleviate spectrum shortage concerns and to address the network congestion issues.

The proposed algorithms exploit an SDN controller's global view of the network to achieve the objectives mentioned above while taking network conditions and the end-user QoS requirements into consideration. With an overall view of the network, the SDN controller has visibility over the offloaded data and can deliver subscribed content from vendors even after offloading, which is important for the operators [36]. SDN-based solutions thus facilitate mobile data offloading by providing centralized coordination and reliable control over the heterogeneous infrastructures. Moreover, the controller's global view of the network provides an ideal platform to design optimal load balancing algorithms. Our contributions also include quantifying the delay incurred due to the SDN-based data processing and forwarding and analyzing the performance improvements of the proposed algorithms under realistic network model. Results show that SDN-based data offloading algorithm decreases threshold miss probability and saves a significant amount of cellular resources simultaneously. It is also shown that SDN-based load balancing achieves better traffic allocation with higher network throughput and a smaller number of handovers, as compared to the baseline algorithms.

The remainder of this chapter is organized as follows: Section 3.2 outlines the network model considered in this work. SDN-based data offloading and load balancing schemes are presented in Section 3.3. The performance of SDN and the proposed algorithms is analyzed in Section 3.4. Simulation results are shown in Section 3.5 while the conclusions are drawn in Section 3.6.

3.2 System Model

We consider a HetNet environment consisting of cellular base stations (BSs) and Wifi access points (APs), as shown in Fig. 3.1. The BSs communicate over the licensed band while Wifi APs communicate over the unlicensed band. The Wi-Fi APs are located randomly within each cell. It is assumed that all of the base stations and the Wi-Fi APs are connected to OpenFlow switches, as depicted in Fig. 3.1. For the cellular network, the switches are co-located with the BSs, i.e., each BS has its OpenFlow-enabled switch. While the switches of the Macrocells are connected to the core network, the Femtocell switches are connected directly to the Internet. On the other hand, for the Wi-Fi APs, the switches are located within the Internet service

provider infrastructure, and each switch can be connected to multiple Wifi APs. All of the switches are controlled by a centralized SDN controller. Given the fact that the controller is only a program running on a server, it can be placed anywhere on the network, even in a remote data center [28].

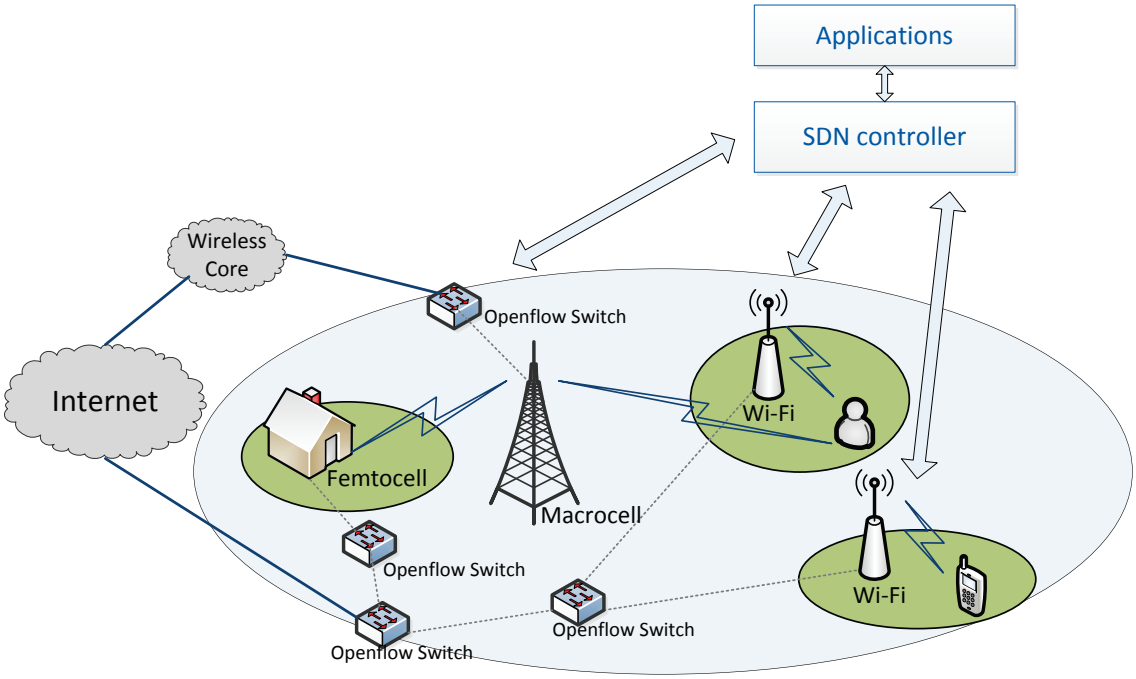

Figure 3.1: SDN-based wireless heterogeneous network structure with control plane design.

Since the SDN controller is a fundamental component of the proposed architecture, it is imperative to discuss the SDN framework in detail. The controller application consists of three core modules, namely, authentication and charging (AC) module, offload manager (OM) and the load balancing (LB) module. The AC includes the identity management and charging record generation functionalities, which are used to enforce admission control and subscription-based charging, respectively. Once authenticated by the AC, a mobile user can access all of the available network resources. These resources are either owned by the network operator or are leased from the available Wi-Fi networks. Next, for the OM module, the service-level rules describe the traffic features and characteristics which are required by the offload manager, e.g., related IP addresses, port numbers, bit rates, and delay sensitivity. These features and characteristics are based on the traffic flow template (TFT) filter [58] and the related QoS descriptions. Finally, the LB module includes the load measurement and mo-

bility management functionalities, which collect cell load ratio and execute the load balancing algorithms. The complete SDN framework is presented in Figure 3.2, where it can be seen that the SDN applications (AC, OM, and LB) utilize global view of the APs to improve network management.

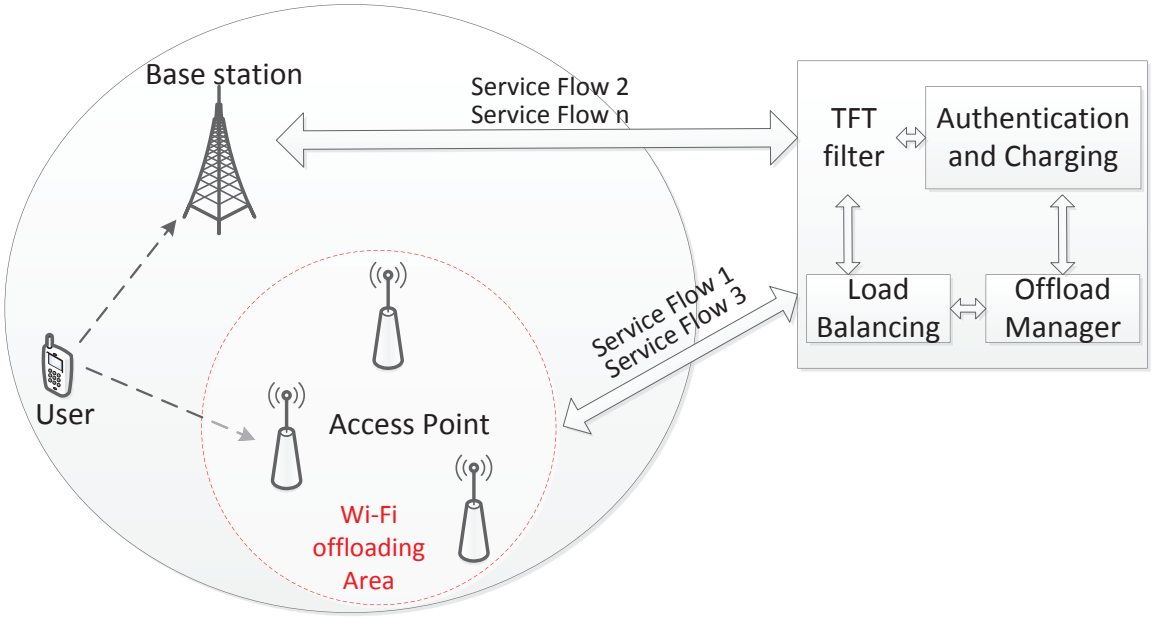

Figure 3.2: SDN-based offloading scenario and module framework.

3.3 SDN-based Traffic Management

In this section, we present SDN-based traffic management algorithms for the heterogeneous network shown in Figure 3.1. To this end, we attempt to alleviate spectrum shortage and network congestion in the cellular network by using SDN-enabled mobile data offloading and load balancing. More specifically, to address the shortage of spectrum, data from the cellular network is offloaded onto a Wi-Fi network whenever the cellular users move within the Wi-Fi range. On the other hand, to deal with network congestion, load balancing is utilized to distribute traffic across multiple cells evenly. For both these cases, all of the decision makings is accomplished by the centralized SDN controller. As mentioned previously, the introduction of SDN facilitates efficient coordination between the cellular and Wi-Fi networks. Moreover, the performance of load balancing can also be improved significantly by exploiting the controller's

view of the whole network. In the next subsection, we present an SDN-based data offloading algorithm.

3.3.1 SDN-based Partial Data Offloading Algorithm

In this subsection, we present an SDN-based partial data offloading algorithm. Here, partial offloading refers to the fact that only part of the user data is offloaded onto the Wi-Fi network, while the remaining traffic is transferred across the cellular network. To elaborate further, if the Wi-Fi network is unable to meet the requirements of the delay and loss sensitive flows, OM only offloads a limited amount of data. As we show later through analysis and simulations, the proposed algorithm improves application performance by taking various service requirements into consideration.

The proposed algorithm proceeds as follows: when a mobile user moves within the Wi-Fi coverage area, and the data file has finished queuing, the OM executes a delay threshold-based selection algorithm, which selects the appropriate traffic flows for partial offloading. Because Wi-Fi provides high data rates without QoS guarantees in the unlicensed band, the choice of offloading traffic should be discreet. Our goal here is to reduce cellular usage by leveraging Wi-Fi connectivity when available but to do so without affecting application performance [38]. The algorithm used to select the appropriate flows for offloading works as explained below:

Step 1) The OM collects the traffic flow requirements of all the users within the Wi-Fi offloading area and calculates the resources required by each user. Assuming each user has a traffic demand u_i and a link rate r_i^w to the Wi-Fi AP, every Wi-Fi user's resource demand is computed as $\theta_i^w = u_i / r_i^w$ [59].

Step 2) The OM calculates the amount of data that can be transferred within the delay tolerance threshold over the Wi-Fi network. If Wi-Fi cannot transfer all of the data before the deadline, partial offloading is executed. Otherwise, all of the data is offloaded onto the Wi-Fi network. Next, we calculate the application specific delay tolerance threshold, T_s, and the amount of data, V_s, that can be transferred within T_s over the Wi-Fi network.

Step 3) OM updates the record of available Wi-Fi resources and the link rate r_i^w that Wi-Fi can provide. This record is maintained for a specific amount of time, and it enables rapid

decision making when a new traffic flow arrives.

3.3.1.1 Calculation of T_s

Assume that a mobile user enters and leaves a Wi-Fi offloading area at times t_{in} and t_{out}, respectively, as shown in Figure 3.3. Consequently, the Wi-Fi connection time is $t_c = t_{out} - t_{in}$. Figure 3.3 also shows the residual time of t_c, i.e., t_r, which is the duration of data transfer from t. In this scenario, the delay tolerance threshold T_s is equal to the difference between the application specific deadline T_d and the waiting time during SDN processing, denoted by D. The calculation of D will be explained later in the next section.

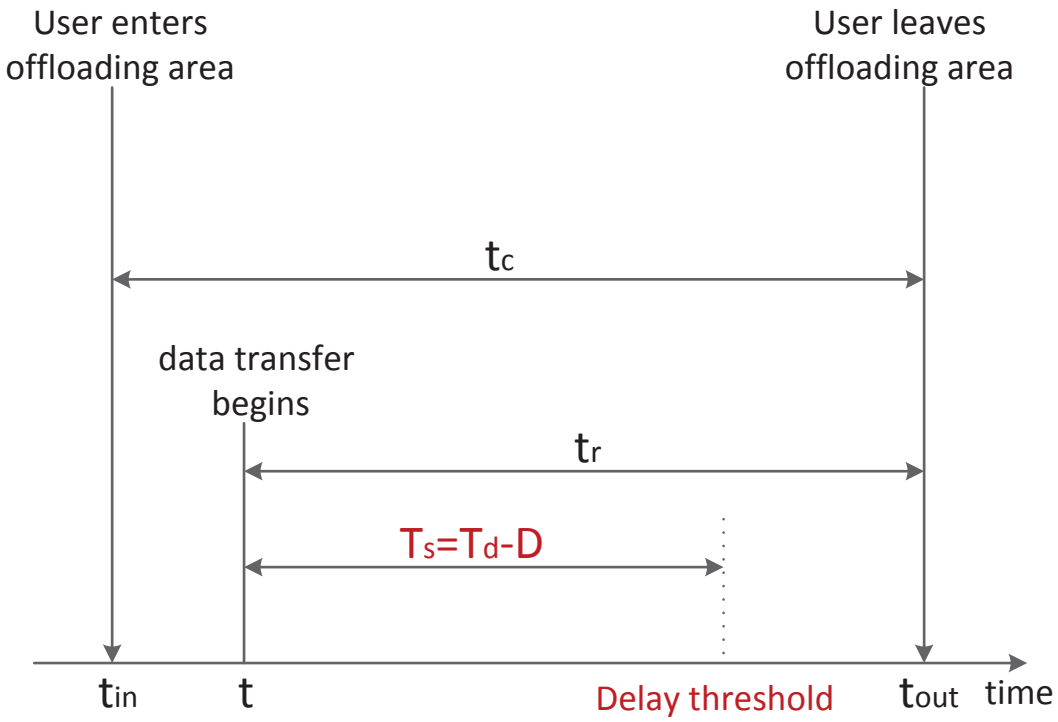

Figure 3.3: SDN-based offloading time diagram model.

3.3.1.2 Calculation of V_s

Let b_1 and b_2, respectively, denote the bandwidths allocated by the cellular network (primary resource) and the Wi-Fi network (secondary resource). As shown in [39], V_s can be written as:

$$V_s = b_1 T_s + b_2 \, min(t_r, T_s). \tag{3.1}$$

Based on the value of V_s, OM decides between partial offload and total offload as shown in Algorithm 1:

Algorithm 1 Partial Data Offloading Algorithm

Delay threshold:T_s
Total file size:$d = u_i$
Size in bytes to be transferred in Wi-Fi within T_s:$V_{s2} = b_2\ min(t_r, T_s)$
if $d < V_{s2}$ **then**
    offload all data to Wi-Fi and update d
  if $d \geq V_{s2}$ and $d_2 = d - V_{s2}$ **then**
    send V_{s2} *on Wi-Fi,* d_2 *on cellular concurrently and update d*
  end if
end if

Here the percentage of b_1 and b_2 are specified in next section.

3.3.2 SDN-based Load Balancing Mechanism

Load balancing [45] can reduce network congestion over an area by distributing user traffic across neighboring APs or BSs. With load balancing, a proportionate share of wireless traffic can be guaranteed for better resource utilization. Since it is hard to guarantee QoS with Wi-Fi, load balancing and vertical handovers of the edge users are seen as the enabling solutions to tackle network congestion. Both these strategies improve QoS by equally distributing the traffic load across the network [36].

Although load balancing has its advantages, yet, there are scenarios where it can prove to be counter-productive, e.g., in low-latency applications such as voice or live (unbuffered) video streaming. More specifically, for mobile users, a high number of handovers impacts the voice quality, and it makes the streaming video jittery. With SDN-based load balancing, however, one can take advantage of the controller's view of the whole network. This makes it easier to find the optimal neighboring cell for load balancing with minimum handovers. SDN-based load balancing proceeds as follows:

Step 1) When a source cell i becomes overloaded, it sends a request to the controller, enquiring about the target neighboring cell for load balancing. A cell i is identified as overloaded when its load ratio LR_i exceeds a certain threshold. The load ratio LR_i is formulated as [45]:

$$LR_i = w_1 * U_i + w_2 * R_i, \tag{3.2}$$

where U_i is the proportion of the number of UEs to cell i's maximum UE capacity, R_i is the ratio of the used resource blocks to the total resource blocks in a cell i, while w_1 and w_2 represent the weight parameters which provide the operators with the option to give higher preference to either U_i or R_i.

Step 2) The LB module calculates all of the neighboring cells' environment state ES_j, which is the average load state of each neighbor cell j's adjacent cells, excluding cell i (denoted as Layer 2 in Figure 3.4).

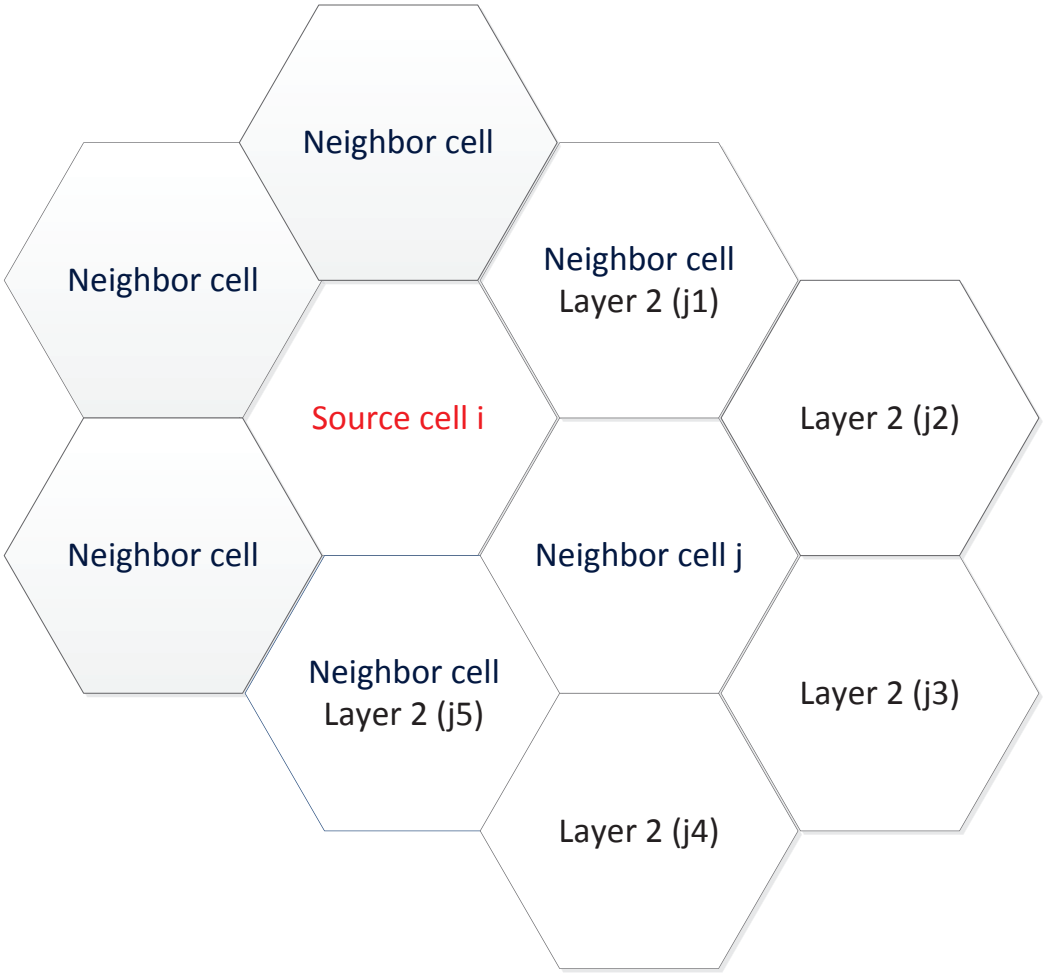

Figure 3.4: SDN based Load balancing with network view.

The environment state of the neighboring cell j can be written as

$$ES_j = (LR_{1j} + LR_{2j} + ... + LR_{nj})/n, \tag{3.3}$$

where n is the number of the neighbor cell j's second layer cells.

Step 3) The LB module computes the overall state OS_j of each neighboring cell of the source cell i and selects the target cell with smallest overall state value for load balancing.

The overall state of each neighboring cell j is a combination of its own load and its environment [60], which is computed as:

$$OS_j = \mu LR_j + (1 - \mu)ES_j, \tag{3.4}$$

where μ is the influence degree of the neighbor's load and its environment.

In the traditional distributed load balancing scenarios, there is a possibility that multiple overloaded cells choose the same target cell for load balancing, causing a new overload situation. Distributed solution thus takes more rounds to achieve an optimized state. However, SDN-based load balancing mechanism uses an overall network view when selecting the target cell, which decreases handover times and improves system performance.

Moreover, by maintaining a list of BSs, SDN-based load balancing encourages new clients to associate with the least loaded BS upon arrival. Then, after a certain amount of time, if there exists an overloaded cell, the controller uses load balancing mechanism to handover appropriate cell edge users. Along with partial data offloading algorithm, it is expected that SDN-based module framework will achieve an optimized overall system state.

3.4 Performance Evaluation

In this section, we analyze the performance of the traffic management schemes introduced in the previous section. To this end, we begin by formulating the delay incurred due to the processing at the SDN-based controller and switches.

3.4.1 SDN Network Delay D

The average delay introduced by SDN depends on the state of the flow table within the switch, i.e., whether or not the switch's flow table contains a rule for the incoming traffic flow. Figure 3.5 shows the queuing model for the controller and the switch.

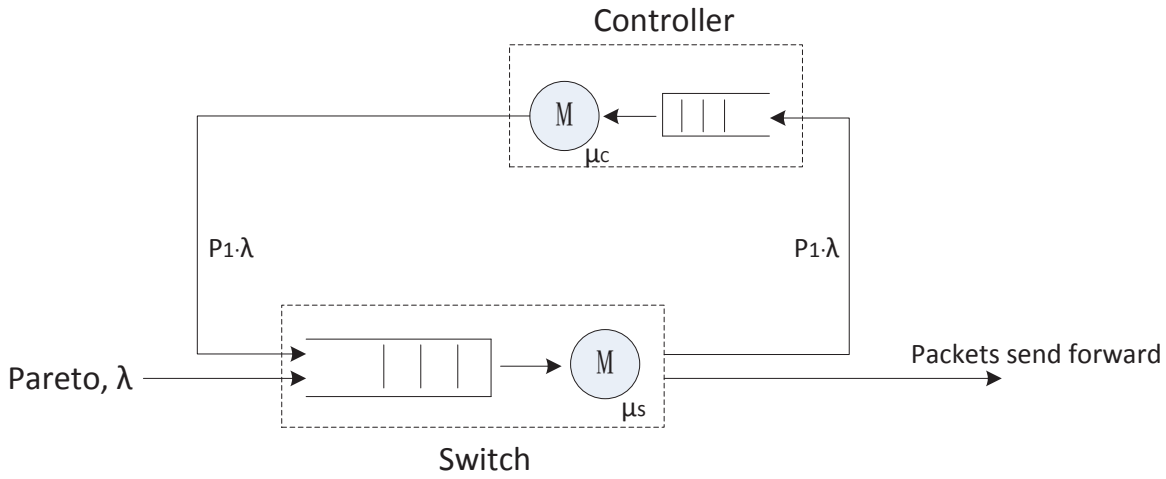

Figure 3.5: A model for SDN switch and controller.

In Figure 3.5, λ is the data arrival rate while μ_s and μ_C represent the processing rates at the switch and the controller, respectively. As shown in the figure, if the packet arriving at the switch is the first packet of a new data flow (new source-destination pair), the switch forwards this packet to the controller. The controller decides the optimal forwarding rule for this packet and returns it to the switch. Moreover, the controller also pushes the corresponding forwarding rule into the flow table at the switch. Subsequent packets from the same flow are forwarded based on the newly installed forwarding rule. It is worth mentioning here that the following analysis is based on the assumption that the incoming traffic is TCP, i.e., a given source node initially transmits a single packet to initiate the TCP handshaking procedure, and the actual data is transmitted once the TCP session is established.

Assuming P_1 represents the probability that there is no flow entry in the OpenFlow switch for the incoming packet, the average packet delay, D, can be written as

$$D = D_1 \times P_1 + D_2 \times (1 - P_1), \tag{3.5}$$

where D_1 is the delay incurred if the switch has to forward the packet to the controller while D_2 represents the delay which occurs when the switch forwards the data directly, i.e., a forwarding rule for the packet already exists. It has been shown in [61] that in a normal productive network carrying end-user traffic, a switch observes new flows with a probability of 0.04. Hence, we use this value for the probability P_1.

The delay D_1 can be written as

$$D_1 = T_C + 2T_{PROP} + 2T_{sw}, \tag{3.6}$$

where T_{sw} and T_C, respectively, represent the delays at the switch and the controller, including the queuing and the processing delays. Moreover, T_{PROP} denotes the propagation delay between the controller and the switch. Note that T_{sw} in (3.6) is multiplied by two since the packet has to make two passes through the switch. In the first pass, it is forwarded to the controller for further processing and in the second pass, it is forwarded along the data path once a forwarding rule has been established. On the other hand, $2T_{PROP}$ accounts for the propagation delays when the packet is sent to and from the controller. The delay D_2 is equal to T_{sw} only, i.e.,

$$D_2 = T_{sw}. \tag{3.7}$$

Next we derive the delays T_{sw} and T_C. Here, it is imperative to mention that in order to simplify the analysis, it is assumed that the packets returning from the controller have no impact on the queuing delay at the switch. This assumption is reasonable for two reasons: firstly, the probability P_1 is relatively small. Secondly, the size of the data returning from the controller is also very small as compared to the size of the data buffered in the queue, i.e., controller returns only a single packet at a given time instance while the traffic at switch arrives in the form of multiple flows containing a large number of data packets. The validity of this assumption is verified later through simulations in Section 3.5.

$$z = \alpha[\mu(1-z)]^{\alpha}\{\frac{1}{\alpha e^{\mu(1-z)}[\mu(1-z)]^{\alpha}} - \frac{1}{\alpha(\alpha-1)e^{\mu(1-z)}[\mu(1-z)]^{\alpha-1}} + \frac{1}{\alpha(\alpha-1)(\alpha-2)e^{\mu(1-z)}[\mu(1-z)]^{\alpha-2}} - \frac{1}{\alpha(\alpha-1)(\alpha-2)}[\Gamma(-\alpha+3) - \Gamma^{*}(-\alpha+3, \mu(1-z))]\} \tag{3.9}$$

3.4.1.1 T_{sw}

In order to derive T_{sw}, the inter arrival process at the switch is modeled with Pareto distribution [62, 63], with shape parameter α and scale parameter k [64]. Furthermore, the switch processing times are modeled as Poisson distributed with rate parameter μ_c. With the aforementioned assumption and the $Pareto/M/1$ queuing model, the switch waiting time T_{sw} is given as [63]

$$T_{sw} = \frac{1}{\mu_s(1-z)}, \tag{3.8}$$

where z is the root of the Laplace Transform of the inter-arrival time distribution function. The root z is given by (3.9) shown on top of the next page [65]. In (3.9), $\Gamma^{*}(-\alpha+3, \mu(1-z)) = \Gamma(-\alpha+3)P(-\alpha+3, \mu(1-z))$, where $P(-\alpha+3, \mu(1-z))$ is the incomplete gamma function [65].

3.4.1.2 T_C

The arrival rate at the controller is Poisson distributed since the departure rate at the switch is Poisson [66]. Therefore, T_C can be calculated using the waiting time equation for $M/M/1$ queuing as

$$T_C = \frac{1}{\mu_c(1-\rho_C)}. \tag{3.10}$$

In the above equation, ρ_C represents the controller utilization and it is equal to $P_1 \cdot \lambda/\mu_c$ (See Figure 3.5).

Using (3.6)-(3.10) in (3.5), one can find the average packet delay with SDN-based data forwarding.

3.4.2 Performance Analysis of SDN-based PDO algorithm

In [67], Xiaolong Li *et al.* prove that at the application level, the size of most of the web traffic, including multimedia files and internet documents, is a combination of long-tailed distribution process and forms Pareto distribution. Therefore, this thesis considers the data file d to be transmitted in the offloading session with Pareto distribution. Resultantly, the cumulative distribution function $F_d(X)$, which is the probability that d is smaller than some number X, is formulated as:

$$F_d(X) = 1 - (\frac{k}{X})^{\alpha} \; for X \geq k \tag{3.11}$$

In order to analyze the performance of the partial data offloading scheme, we use two key performance indicators: The first indicator is the delay threshold miss probability, P_{miss}, which is the probability that the Wi-Fi network is unable to meet the service latency requirements; the second indicator is amount of data d_1 that is offloaded from the cellular network. The probability P_{miss} is given as

$$P_{miss} = Pr[d > V_s]. \tag{3.12}$$

Using the cdf of d from (3.11) and the value of V_s from (3.1), the above equation can be written as [39]

$$\begin{aligned} P_{miss} = & \int_{t=0}^{T_s} [1 - F_d(b_1 T_s + b_2 t)] r_c(t) dt \\ & + [1 - F_d((b_1 + b_2) T_s)] \int_{t=T_s}^{\infty} r_c(t) dt, \end{aligned} \tag{3.13}$$

where $r_c(.)$ is the probability density function of the residual time t_r. From equation (3.13), it can be seen that the probability P_{miss} consists of two parts: the first part reflects the scenario where the time t that a user spends in the Wi-Fi connection area is smaller than the application deadline time T_s. In this situation, the transmission volume in the Wi-Fi network is $b_2 t$. On the other hand, the second part of the equation addresses the scenario where the duration of a user's stay in the Wi-Fi area is larger than the application deadline time. Accordingly, the transmission volume of Wi-Fi network in this situation is equal to $b_2 T_s$.

By setting b_1 in (3.13) equal to 0, one gets P_{miss} for the scenario when all of the traffic is offloaded onto the Wi-Fi network. Recall that t_r is the residual life of t_c. According to the

distribution of the asymptotic residual life from the renewal theory [68], the probability density function $r_c(.)$ can be formulated as:

$$r_c(t_r) = \frac{1 - P(T < t_r)}{E[t_c]} = \frac{1 - F_c(t_r)}{E[t_c]}, \tag{3.14}$$

where $E[t_c]$ is the expected value of the connection time t_c while $F_c(.)$ is the distribution function of t_c.

Assuming that t_c follows Erlang distribution with parameters (n, λ_e), which is a widely used traffic model, the distribution function $F_c(.)$ of connection time t_c can be formulated as [68]

$$F_c(t_c) = \frac{\gamma(n, \lambda_e t_c)}{(n-1)!} = 1 - \sum_{m=0}^{n-1} \frac{1}{m!} e^{-\lambda_e t_c} (\lambda_e t_c)^m, \tag{3.15}$$

where $\gamma(.)$ is the lower incomplete gamma function and $E[t_c]$ is equal to n/λ_e. By using (3.14) and (3.15) in (3.13), one obtains P_{miss} as

$$\begin{aligned} P_{miss} = (\frac{\lambda_e}{n}) \sum_{m=0}^{n-1} \int_{t=0}^{T_s} (\frac{k}{b_1 T_s + b_2 t})^\alpha \frac{e^{-\lambda_e t} (\lambda_e t)^m}{m!} dt + \\ (\frac{k}{(b_1 + b_2) T_s})^\alpha \{1 - (\frac{1}{n}) \sum_{m=0}^{n-1} [1 - \sum_{j=0}^{m} \frac{e^{-\lambda_e T_s} (\lambda_e T_s)^j}{j!}]\}, \end{aligned} \tag{3.16}$$

where $T_s = T_d - D$. P_{miss} indicates the performance of partial offloading scheme as a function of the delay tolerance threshold T_s. One can find the improvement of proposed algorithm by comparing P_{miss} under various primary bandwidths b_1.

The second indicator, the amount of offloaded data, i.e., d_1, which is served by the Wi-Fi network can be expressed as:

$$d_1 = \begin{cases} \frac{b_2}{b_1 + b_2} d, if\ d > b_2 t_r \\ b_2 t_r, otherwise, \end{cases} \tag{3.17}$$

which means that if data volume is larger than Wi-Fi capability, the data is transmitted concurrently in cellular and Wi-Fi networks. Otherwise all the data is offloaded onto the Wi-Fi

network. Equation (3.17) can be re-formulated as

$$d_1 = \frac{b_2 d}{b_1 + b_2} Pr[d > b_2 t_r] + b_2 t_r Pr[d \leq b_2 t_r] \tag{3.18}$$

Using a similar procedure as that presented in [39], the above equation can be written as

$$d_1 = \frac{b_2 d}{b_1 + b_2}[1 - R_c(\frac{d}{b_2})] + b_2 \int_0^{d/b_2} t_r r_c(t_r) dt_r, \tag{3.19}$$

where $R_c(.)$ is the distribution function of t_r and $1 - R_c(\frac{d}{b_2})$ shows the probability that the value $\frac{d}{b_2}$ is larger than time value t_r. Using (3.14), one obtains $R_c(.)$ as

$$\begin{aligned} R_c(t_r) &= \int_{t=0}^{t_r} r_c(t) dt \\ &= (\frac{1}{n}) \sum_{m=0}^{n-1} [1 - \sum_{j=0}^{m} \frac{e^{-\lambda_e t_r} (\lambda_e t_r)^j}{j!}] \end{aligned} \tag{3.20}$$

Finally, by replacing t_r in the above equation with $\frac{d}{b_2}$ and using the resulting equation in (3.19), one gets the offloaded data volume d_1 as

$$\begin{aligned} d_1 = \frac{b_2 d}{b_1 + b_2} \{1 - (\frac{1}{n}) \sum_{m=0}^{n-1} [1 - \sum_{j=0}^{m} \frac{e^{-\lambda_e x} (\lambda_e x)^j}{j!}]\} \\ + \frac{b_2}{\lambda_e n} \sum_{m=0}^{n-1} (m+1)[1 - \sum_{j=0}^{m+1} \frac{e^{-\lambda_e x} (\lambda_e x)^j}{j!}], \end{aligned} \tag{3.21}$$

where $x = d/b_2$. In particular, $\frac{d_1}{E[d]}$ describes the proportion of the data that has been offloaded onto Wi-Fi network.

3.4.3 Performance Analysis of SDN-based Load Balancing Algorithm

The performance of SDN-based load balancing mechanism is measured by the number of handover times, the equilibrium extent of the network and the throughput of the network.

Equilibrium extent is defined as the degree of load balanced across the entire network [45]

and it can be written as

$$\nabla(t) = \frac{(\sum_c \rho_c)^2}{|N| \sum_c (\rho_c)^2}, \tag{3.22}$$

where N is the number of cells and ρ_c is the load density of cell c. Obviously, the network resource is better utilized if the load is more evenly balanced across the network. In this thesis, we use the over throughput of the network to measure the performance of the network resource utilization.

3.5 Performance Evaluation

In this section, we evaluate the performance of SDN and the proposed algorithms. To this end, we begin by evaluating the performance of SDN-enabled switches and controller.

3.5.1 Performance Evaluation of SDN

In this section, the performance of the SDN framework, shown in Figure 3.5, is evaluated in terms of network utilization and the incurred delay. For the simulation setup, the service time of controller and switch was set to be $0.33ms$ and $9.8\mu s$, respectively [66, 69]. To account for different types of data traffic, we used different values for the shape parameter α of the Pareto arrivals at the switch. More specifically, values of $\alpha = 1.5$ and $\alpha = 2.5$ were used. Moreover, we also evaluate the performance when the arrival process at the switch is Poisson distributed. All the simulation results were averaged to 10000 number of iterations.

Figure 3.6 shows the SDN delay as a function of the network utilization ρ. Recall that in the queuing model introduced in Section 3.4.1, it was assumed that the packets returning from the controller to the switch cause negligible delay in the switch queue. To justify this assumption, in Figure 3.6, we also plot the SDN delay which occurs when the effect of the packets returning from the controller is not ignored (Model 2 in Figure 3.6). From the figure, it can be seen that the simulation results with and without the aforementioned assumption follow approximately the same trend and therefore, our assumption is justified. Moreover, it can also be seen that the theoretical and the simulation results also match very closely with each other, thus verifying the validity of the SDN processing delay analysis in Section 3.4.1.

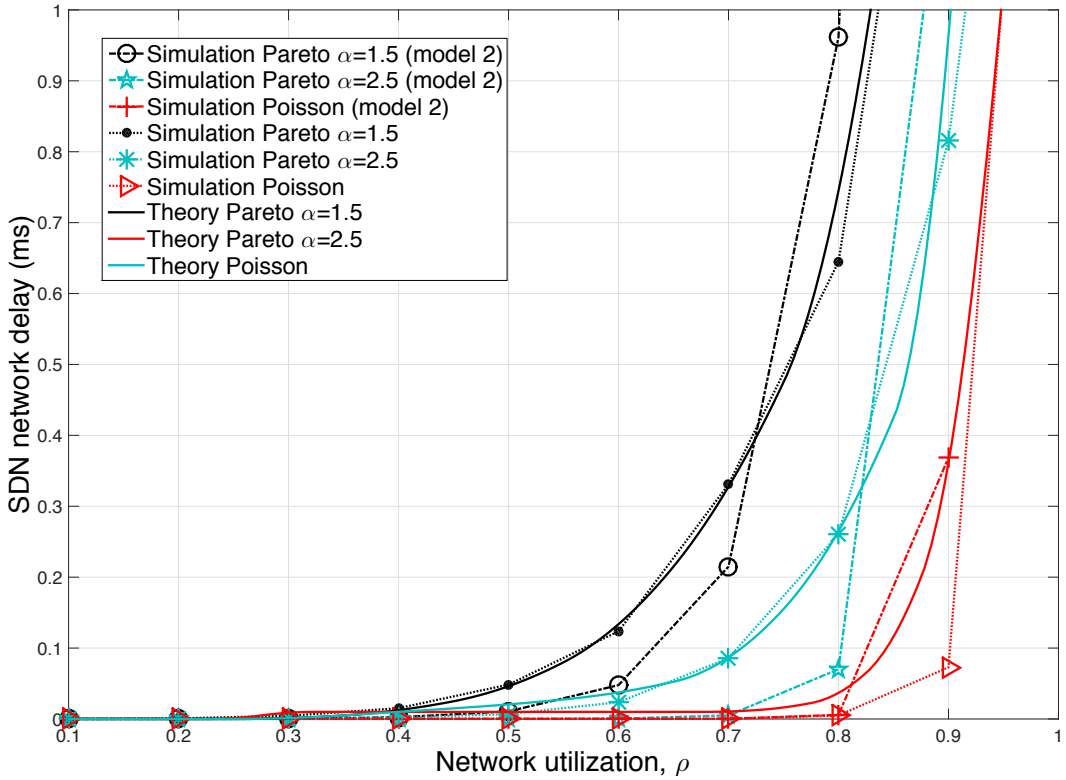

Figure 3.6: SDN network delay versus network utilization using different SDN queuing model.

As mentioned previously, different α values represent different types of application traffic [3]. It can be seen that with decreasing α, the network delay increases under the same utilization rate, which shows that the smaller traffic flows cause greater burden on the SDN network due to the higher arrival rate. Traditional telecommunication voice traffic, shown by Poisson arrival, has the lowest delay since it does not participate in the SDN offloading procedure. Based on the above discussion, it can be concluded that SDN-based solution is more suitable for applications which are less sensitive to latency.

3.5.2 Performance Evaluation of SDN-based Partial Data Offloading

In this section, we evaluate the performance of the proposed partial data offloading scheme. To this end, all the simulations were conducted in Matlab and for each simulation round, a user randomly moved across the offloading area with residence time t_c, which followed an Erlang distribution [39]. A download traffic of size d was initiated at a given time instance t whenever a user resided within the offloading area. For demonstration purposes, it was assumed that the

session data traffic d had a shape parameter of $\alpha = 1.5$ and scale parameter of $k = E[d](\alpha-1)/\alpha$ [64]. The default Wi-Fi bandwidth, b_2, was set to 5 Mb/s.

Figure 3.7 shows the threshold miss probability P_{miss} as a function of the delay threshold, T_s in seconds. For this simulation, it is assumed that the residence time t_c is 30 min while the average data size $E[d]$ is 10 MB. The partial offloading algorithm is used to allocate different amounts of primary bandwidth, b_1. When $b_1 = 0$, the data is offloaded completely onto the Wi-Fi network. Figure 3.7 shows that when the delay threshold is small, especially from $5 \sim 20$ seconds, partially offloading the data can effectively improve the threshold assurance. That is to say, if an application is sensitive to delay, the use of partial offloading improves the application performance by $20\% \sim 50\%$. Finally, it can also be seen from the figure that if the application is not sensitive to delay, e.g., delay threshold is higher than 20 seconds, the offloading performance in terms of threshold miss probability becomes constant for a given primary bandwidth.

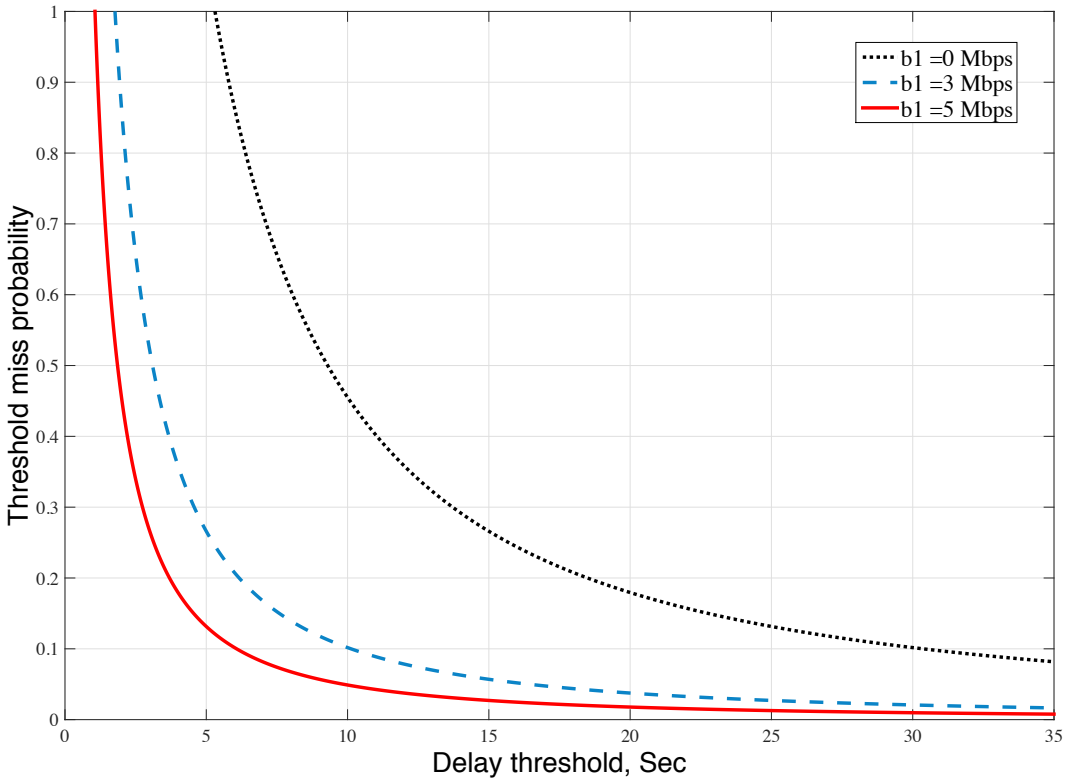

Figure 3.7: Performance of partial offloading algorithm in threshold miss probability versus delay threshold.

Figure 3.8 plots the amount of offloaded data as a function of the average residence time,

t_r. It is obvious from the figure that when the average residence time is larger than 6 min, the amount offloaded data remains unchanged. The reason is that although the session offloads more data if the user stays longer in the offloading area, the maximum amount of offloaded data is limited by $E[d]$. Figure 3.8 shows that the amount of offloaded data increases significantly with increasing $E[d]$. This is reasonable because when a cellular network has higher loads, offloading will play a more significant role.

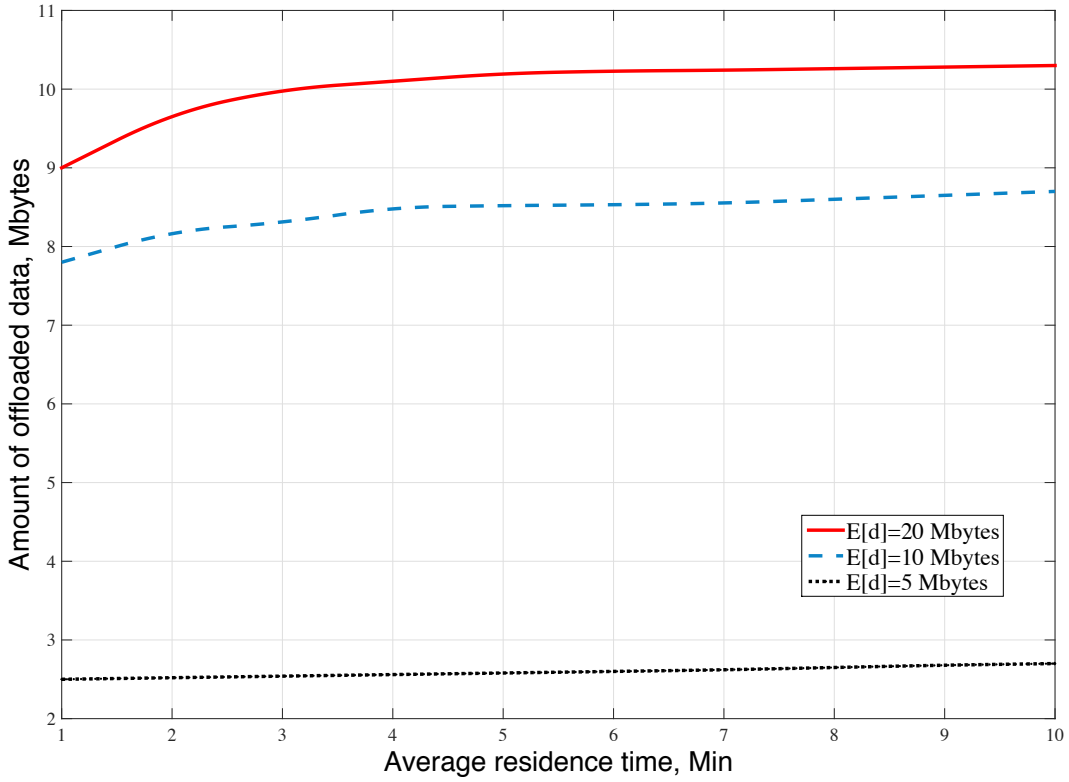

Figure 3.8: Offloaded data volume by Wi-Fi data offloading in terms of different arriving traffic volume.

3.5.3 Performance Evaluation of SDN-based Load Balancing

For the simulation of proposed SDN-based load balancing mechanism, we set up two reference scenarios: traditional load balancing and distributed mobility load balancing (DLB) [45]. In traditional load balancing, the decisions are made based on the received signal strengths. On the other hand, in DLB, the handover parameter is adjusted dynamically according to cell load measurement $LOAD_c = min(\frac{\sum_{u \in c} N_u}{N_{tot}}, 1)$ [45], which is the ratio of required resources, N_u, of all

the users u to the total number of resources, N_{tot}, in the cell c.

Figure 3.9 depicts the number of handover instances as a function of the simulation runtime. A lower number of handovers is desirable since frequent handovers affect service quality and user experience. It can be seen from the figure that SDN-based LB has a fewer number of handovers as compared to the other two scenarios. This happens because with the better knowledge of all the cell load states and trends, SDN-based LB is able to select the most suitable target cells more efficiently while DLB takes more rounds, consequently requires more time to achieve the optimized state.

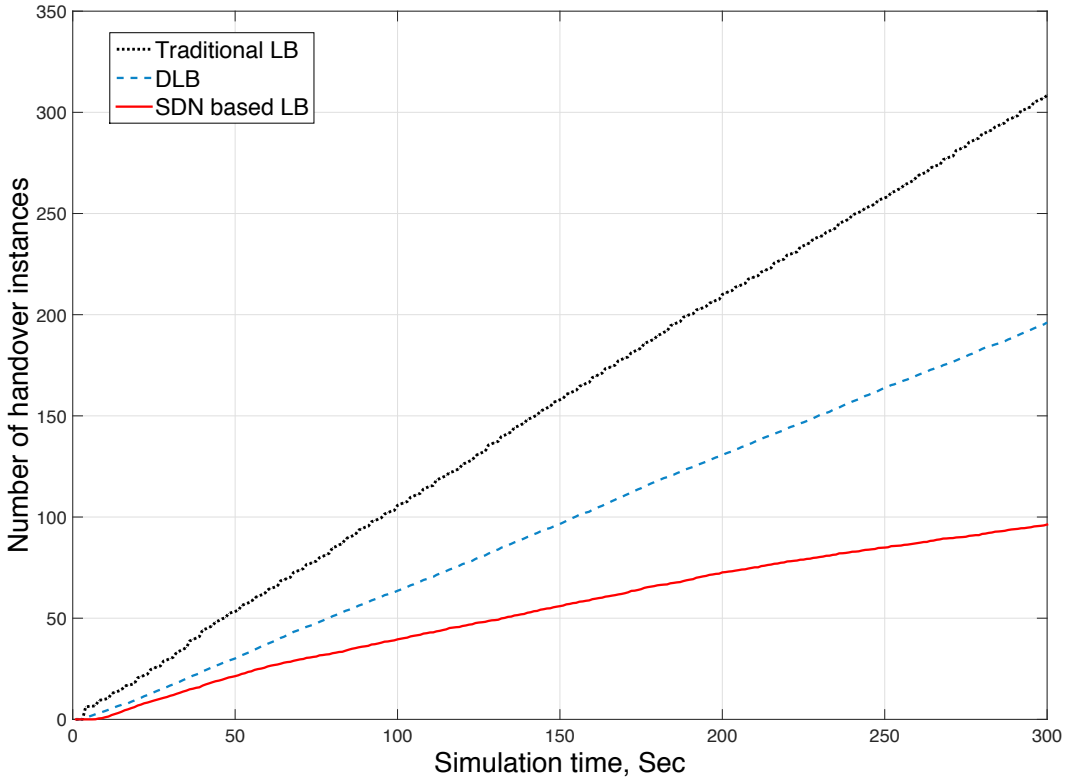

Figure 3.9: Cumulative handover times as a function of the simulation time.

Figure 3.10 shows the extent of equilibrium between cells as a function of the simulation runtime. It can be seen that with increasing simulation time, traffic load becomes more balanced with all the LB mechanisms. However, SDN-based LB outperforms the baseline methods. This is because the baseline methods only have a limited view of the network. Notice that at the start of the simulation, SDN-based LB performs worse than the other two methods. This happens because the baseline methods choose the neighboring cell with lowest load as the tar-

get cell for LB, which provides faster balancing. However, the target cell eventually becomes overloaded, if its surrounding environment is mostly highly loaded. On the other hand, with an overall view of the network, SDN-based LB uses the overall state to select the target cell and thus guarantees a long-term advantage in equilibrium extent.

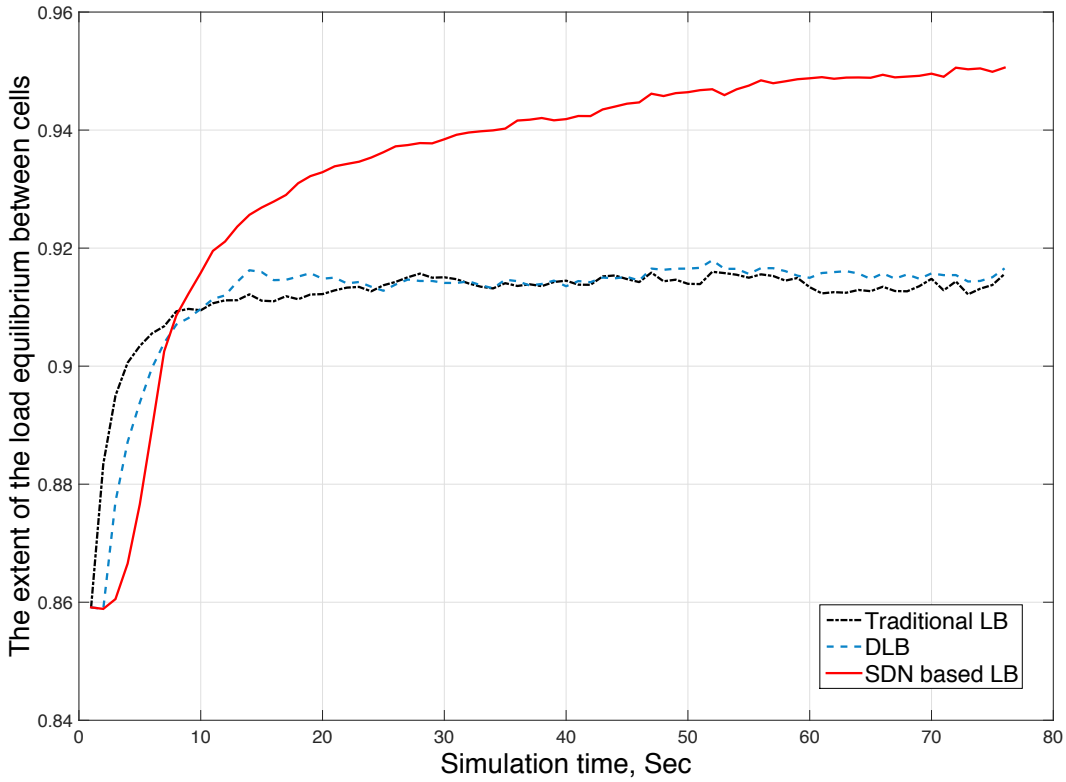

Figure 3.10: The extent of equilibrium as a function of the simulation time.

The load balancing performance can be further verified by comparison of the network throughput, as shown in Figure 3.11. Obviously, the trend of throughput is similar to the load balancing extent in Figure 3.10. This is easy to understand: when the load is more balanced within the network, the resources are utilized more efficiently, rendering a higher throughput for the whole network.

3.6 Chapter Summary

Due to the increased data traffic and the co-existence of different radio access technologies, efficient traffic management is a key challenge for future 5G networks. In this chapter, we

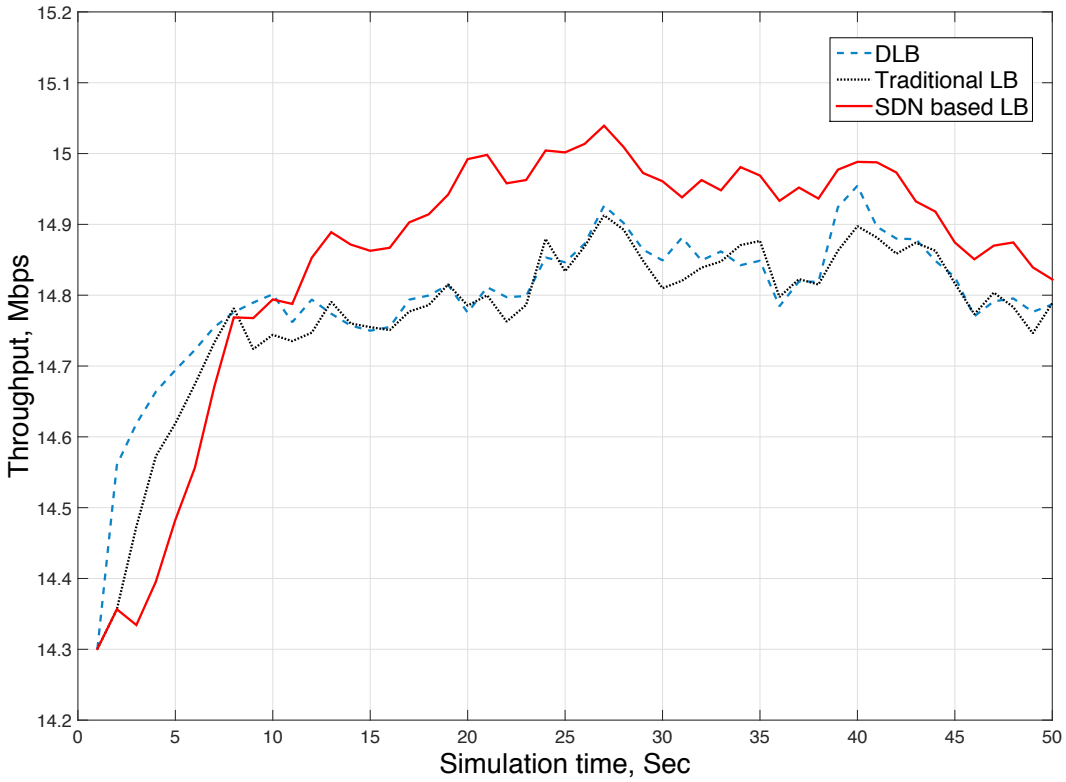

Figure 3.11: Network throughput comparison as a function of the simulation time.

presented SDN-based Wi-Fi data offloading and load balancing algorithms. The new algorithms utilized the controller's global view of the network to take more informed decisions for efficient traffic management. We also analyzed the performance of the proposed algorithms under realistic load conditions. To this end, we first introduced a queuing model with Pareto arrivals to investigate the processing and forwarding delays incurred due to the SDN architecture. Then, we analyzed the performance of the proposed SDN-based partial data offloading scheme regarding the threshold miss probability and the amount of data offloaded successfully onto Wi-Fi. Through simulations, it was shown that partial data offloading saves primary resources and decreases threshold miss probability by 20% $\sim$ 50%, which ultimately improves the application performance at the user end. Furthermore, the simulation results also confirmed that SDN-based LB outperforms the baseline methods by minimizing the number of required handovers by 50% and by balancing the loads more evenly across multiple cells. Our results and discussions showed that the delay incurred by SDN is well within the acceptable limits for most applications. Particularly, it has been demonstrated that SDN-based solutions per-

form better for large data traffic with high delay tolerance. All in all, SDN is proved to be a suitable enabling technology for introducing intelligence within the wireless networks and for providing fine-grained control to the network operators.

Chapter 4

SDN-enabled Traffic Offloading and Beamformed Transmission in 5G-VANET

4.1 Introduction

With the recent advancement of artificial intelligence and sensor technologies, autonomous (i.e., self-driving) vehicles can sense their surroundings in real-time by the combined use of many techniques including radar, lidar, GPS, Odometer, and computer vision. Consequently, the autonomous vehicle is closer to reality than we ever thought. According to the recent market survey published by McKinsey [14], it is envisioned that more than 15 million self-driving cars will be on the road by 2030. We can imagine that very soon drivers would be released from the burden of driving and thus have time for mobile Internet access. As such, handling of the growing vehicle-generated data traffic in cellular-vehicular ad hoc networks (VANET) [70], which involves ad hoc communications among nearby vehicles and between vehicles and nearby roadside equipment, has attracted much attention from both academia and industry.

However, 5G-VANET has inherent challenges in supporting extremely dynamic vehicle-related data traffic. Due to the high mobility of vehicles and their irregular distribution, VANET has dynamic network topology, where vehicles could join or leave the network quickly, and the links between vehicles connect and disconnect very often. As such, timely updating of road traffic topology is essential for the operation of 5G-VANET. These challenges are further

compounded by more stringent latency requirement of 5G [9], hence requiring more consistent link quality. When the densified vehicles all directly communicate with the cellular base station (BS) or roadside units during the rush hour, the high volume of concurrent Vehicle-to-Infrastructure (V2I) communications may lead to extremely high signaling overhead and outage probability. Consequently, adaptive, reliable and situation-aware management of dynamic traffic is critical for the operation of 5G-VANET.

Furthermore, 5G-VANET is also impeded by the heterogeneity and ossified cellular network architecture due to the use of diverse access networks and vendor-specific equipment. Due to the inevitable network densification in the quest for high data rate and the mixed use of different wireless technologies, 5G is envisioned to have a heterogeneous network (HetNet) architecture [71]. The intractable interconnection and the limited information sharing between HetNet infrastructures and different operators bring additional difficulty for vehicle traffic management. Additionally, with the increasing complexity of the future 5G network, the vendor-specific hardware and protocols makes it challenging and remarkably expensive for operators to dynamically adapt their network operations.

In addressing these challenges, we propose a Software-Defined Networking (SDN) enabled 5G-VANET with the capability of adaptive vehicle clustering and beamformed transmission in supporting the aggregated traffic from the cluster head. Through the separation of data plane and control plane [21], SDN enables the 5G-VANET management and facilitates the centralized control over HetNets by providing a global network view and a unified configuration interface despite the underlying heterogeneous networks involved. With its open and reconfigurable interface, SDN provides an enabling platform to apply intelligence and consistent policy for 5G-VANET HetNets. In the proposed 5G-VANET, arriving road traffic will be predicted with the assist of SDN to achieve adaptive vehicle clustering. Within each vehicle cluster, a cluster head (CH) is selected to aggregate traffic from other vehicles and communicate with the cellular BS to reduce signaling overhead. A dual CH design is then proposed to guarantee the robustness and seamless trunk link communication. Furthermore, when there is high capacity demand over the trunk link from intra-cluster vehicles, an adaptive beamformed trunk link transmission scheme and cooperative communication are considered as potential solutions for improving the 5G-VANET communication quality, capacity and reducing the traffic distribu-

tion latency.

The remainder of this chapter is organized as follows: Section 4.2 presents the network architecture of the SDN-enabled 5G-VANET. Based on the network model, SDN-enabled adaptive vehicle clustering, and dual cluster head design are then proposed and elaborated in Section 4.3. The beamformed adaptive transmission scheme and cooperative communication are explained in Section 4.4, while the performance of SDN-enabled 5G-VANET applications is analyzed in Section 4.5. The conclusions are drawn in Section 4.6.

4.2 Overall Network Architecture of SDN-enabled 5G-VANET

The overall network architecture of the proposed SDN-enabled 5G-VANET, which consists of a HetNet environment, as shown in Fig.4.1, is designed to support adaptive vehicle clustering and trunk link traffic aggregation schemes. It features a layered architecture consisting of macrocells and small cells, i.e., the base station (BS) and access point (AP), etc. As shown in this figure, vehicles are moving across the cells, bearing dynamic traffic requirements. To provide a consistent policy and global management of the 5G-VANET resources, macrocell BSs and APs are all controlled by a centralized SDN controller with OpenFlow protocol through high-capacity fiber optic links [21]. SDN removes the control logic from the underlying infrastructures (e.g., BSs and APs) to control layer so that applications can then be implemented on the central SDN controller to provide new functions including adaptive clustering and traffic management over the whole 5G HetNets.

The right side of Fig. 4.1 shows the operational architecture of the proposed 5G-VANET architecture. The controller is responsible for the global policies related to access network, including authentication, mobility/traffic management, clustering scheme and other global issues, while the BSs (APs) constitute the data plane of the SDN-enabled HetNets and implement the controller-defined policies. Next, the two components (i.e., the BSs and the SDN controller) will be discussed in more detail.

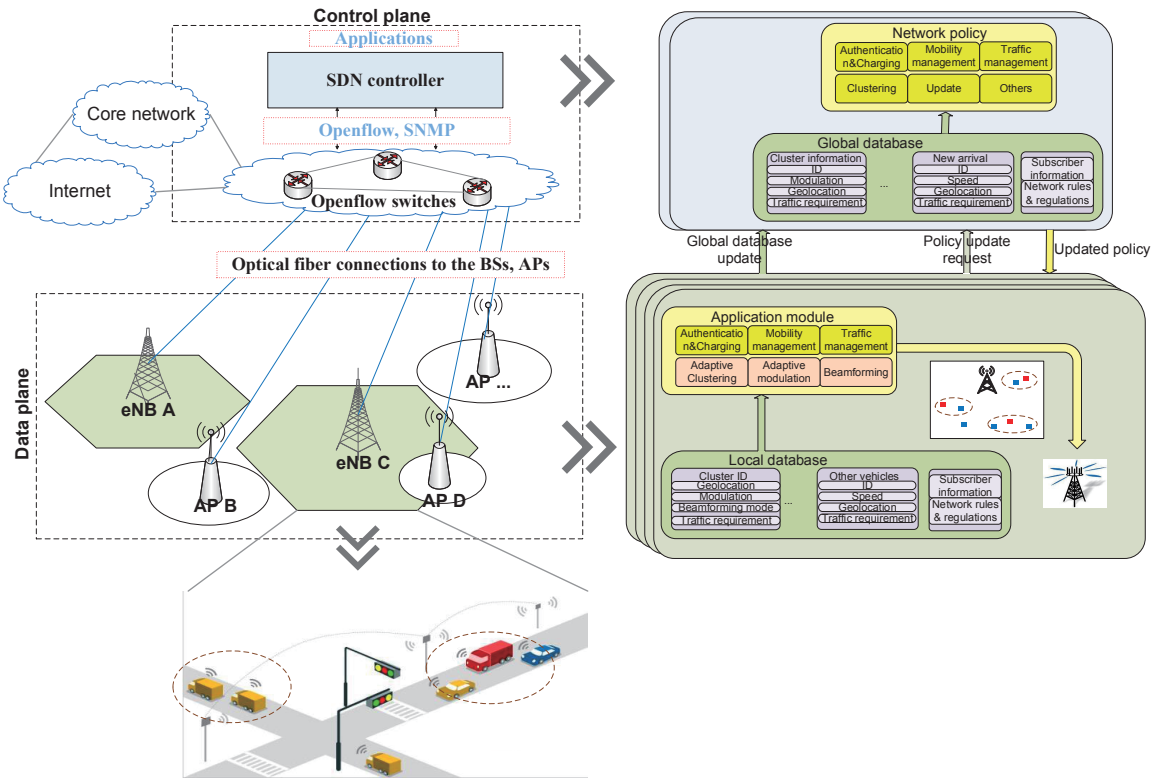

Figure 4.1: SDN-enabled 5G-VANET integrated network architecture and controller-defined network policies.

4.2.1 The Base Stations and Access Points

In the proposed framework for 5G-VANET, each BS or AP has a local database (LDB) and application module. With the support of the LDB, the BS is able to obtain the cell load conditions and facilitates local decision-making [72]. In general, LDB stores the information about vehicles within the cell, including the clustering information, geo-location of vehicles, traffic requirement and transmission scheme. Each of these sections is updated when there is new vehicles access or leave the current cell.

The information gathered from multiple LDB constitutes the global database (GDB), which is then utilized by the SDN controller to design network level policy and update the local application modules, as illustrated in Fig.4.1. Afterward, clusters are formed adaptively, and beamformed transmission scheme of aggregated traffic are determined accordingly. The BSs take care of all local decision-making, e.g., the communication of the other cellular users that are not defined by the centralized controller (such as the communication between pedestrians

and BSs or routine procedures like handover), to reduce the processing burden of the controller.

4.2.2 SDN Controller

SDN controller enables the coordination and information sharing between heterogeneous networks through the separation of control plane and data plane. As shown in Fig. 4.1, the SDN controller includes a global database (GDB) and a network policy-making module. The GDB contains information about all the vehicle clusters of the service area and is regularly updated by the BSs/APs [10]. With the global view over the whole service area, programmable applications can be run on the controller to realize global functions, such as authentication, service charging, adaptive clustering and so on. The updates of the overall network policies could be pro-actively (after a predefined period) or reactively (requested by BS/AP due to cell overloading).

Based on the overall SDN-enabled 5G-VANET network architecture elaborated above, adaptive vehicle clustering scheme and dual cluster head design are proposed in the following section. After the formation of the vehicle clusters, the beamformed adaptive transmission scheme for the trunk link between cluster head and BS is also discussed in detail to support the aggregated traffic from the clustered vehicles.

4.3 Adaptive Clustering in SDN-enabled 5G-VANET

Due to the high mobility of vehicles and the restrictions in their range of motion, vehicle clustering is seen as a promising solution in reducing the overhead of cellular network and providing better communication quality with a low relative speed among clustered vehicles. There have been lots of recent studies in the area of vehicle clustering. Authors in [43] provide a clustering method which considers vehicle velocities, the direction of movement, and inter-vehicle distance. However, it fails to take the unpredictable nature of vehicle mobility into consideration. In [44], a dynamic clustering-based mobile gateway management mechanism is proposed, which considers vehicle mobility and executes clustering algorithm periodically. However, how to decide this period remains an open problem, and the cluster maintenance also dramatically increase the computing burden on cluster head. With a coexistence of multiple

HetNet infrastructures in the future 5G network, it is also difficult for a single base station to predict the arriving traffic and execute clustering algorithms adaptively due to the limited interconnection.

In the proposed SDN-enabled 5G-VANET, the controller's global view over the HetNets and the timely updating of road traffic topology provide a viable solution in addressing the above challenges. As vehicles usually move fast and APs only have limited coverage, we consider that APs only provide updated information of related vehicles and the clustered vehicles would communicate with BSs through a selected vehicle, namely, cluster head (CH). Due to the consistency of moving speed and direction of traveling vehicles, SDN controller will be able to monitor and predict the location of arriving vehicles using different locations and data analytics techniques, and then inform the relevant cellular BSs in advance to guarantee adaptive and efficient clustering, as shown in Fig.4.2. Based on the high level "road topology" collected from heterogeneous BSs and APs of different infrastructures or operators, the proposed clustering algorithm would be executed only when needed instead of periodically. We can also define a traffic threshold and take the delay requirement and size of the upcoming in-vehicle data traffic into consideration when making vehicle clustering decisions.

In Fig. 4.2, vehicles that are moving in two directions are grouped into different clusters. The vehicles that have a cellular interface, i.e., yellow cars in Fig. 4.2, are defined as mobile gateway candidates as they can communicate with cellular networks. A CH is selected from the mobile gateway candidates, and then all other vehicles in the same cluster communicate with BS through the CH. Moreover, communication between CH and other intra-cluster vehicles could be through different wireless protocols, e.g., IEEE 802.11p, to relieve the cellular burden and save licensed spectrum resources. To guarantee seamless communication, a backup CH is also selected from the mobile gateway candidates to record a copy of signaling message, namely, floating car data (FCD), from the existing CH and be prepared for emergencies [73]. Note that in Fig. 4.2, beamforming technique is applied to focus the cellular signal at areas with concentrated vehicles. The vehicle cluster colored blue illustrates the uplink traffic collection procedure, while the cluster colored orange shows the downlink traffic distribution. The cooperative multi-receiver coordinated decoding is used in traffic distribution phase only when the trunk-link traffic volume is higher than a threshold, and the detail will be provided in Section

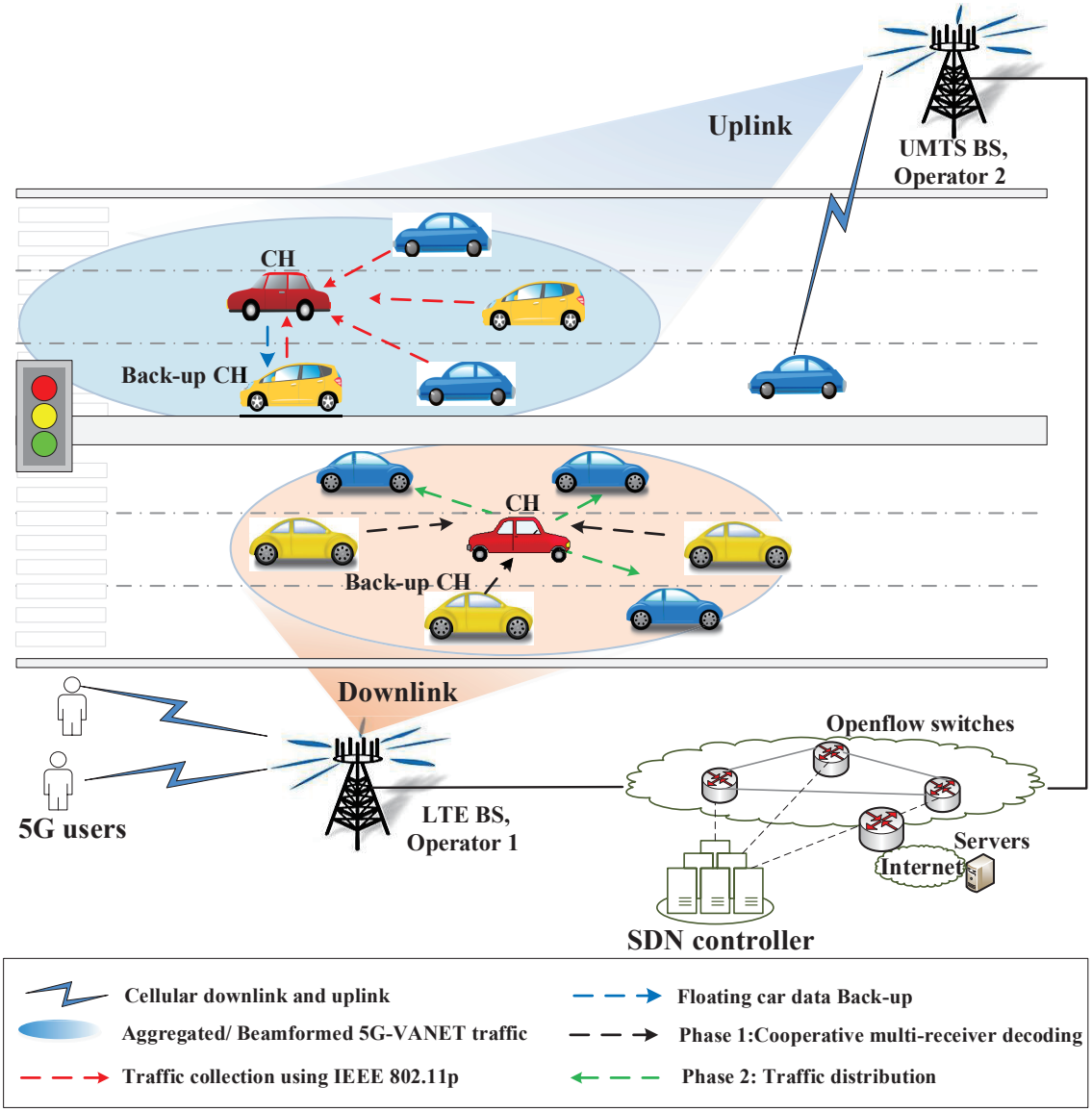

Figure 4.2: SDN-enabled adaptive clustering in 5G-VANET integrated network.

4.4.

Next, we will elaborate the SDN-enabled adaptive vehicle clustering mechanism in 5G-VANET. Specifically, SDN-enabled adaptive clustering is realized under the collaboration of cellular BS and mobile gateway candidates. There are three parameters utilized during the clustering procedure: Angle of arrival (AoA) (θ), received signal strength (RSS), and inter-vehicular distance (IVD). Below, the adaptive clustering procedure is divided into four steps:

1) Base station initialized grouping: With the road traffic topology provided by SDN controller, the BSs are aware of the arriving traffic and prepare themselves in advance. Once

the cell is overload and clustering conditions are met, the vehicles are roughly classified into groups according to similar AoA and RSS, and the member information of each group is sent back to the mobile gateway candidates in the group.

Assume that there are N transmission angles of equal degrees $(360/N)$ in the vehicle transmission surface, and the speed limit of the road is around V_{MAX}, we can define different vehicle groups by the combination of different transmission angle and RSS. Each group is then characterized by

$$\begin{cases} \theta_x - \theta_y \leq \frac{360}{N} \\ RSS_x - RSS_y \leq 1 - e^{-\frac{\Delta V}{a}} \end{cases} \tag{4.1}$$

where x and y represents two vehicles, ΔV is the difference in the mobility speed of two vehicles, and a is a constant that defines the rate of variation of the 5G signal strength when the mobility speed increase or decrease by a unit [44].

2) Vehicle clusters formation: After receiving the vehicle grouping list from the base station, inter-vehicular distance (IVD) would be used by the mobile gateway candidates to refine the group and form the final cluster. As the vehicle position information measured or predicted by base stations might not be accurate, the mobile gateway candidates use broadcasting message (For example, IEEE 802.11p has a transmission range of around 250m) to verify the neighbor vehicles and update the group member list. The inter-vehicle distance of the final clusters is constrained by

$$d \leq R_t \cdot (1 - \varepsilon), \tag{4.2}$$

where R_t denotes the maximum transmission range of IEEE 802.11p protocol and ε reflects the wireless channel fading conditions [43].

3) Dual cluster head selection: After the formation of the clusters, a cluster head would be selected in each cluster in order to effectively relay the vehicle-related traffic to cellular networks. Assume that there are K vehicles in a cluster, the CH selection could be defined by a linear optimization problem as shown below [74]:

$$
\begin{aligned}
& max_{CH_i} \ \ Throu_{CH_i} \\
\text{subject to} \ \ & SNR_i \geq SNR^{min} \\
& V_i - \frac{\sum_{i=1}^{N} V_i}{N} = 0 \\
& V_i \geq 0, \ N \geq 0, \ CH_i \in N
\end{aligned}
\tag{4.3}
$$

The objective of the CH selection is to maximize the throughput rate of trunk link under the constraint of channel quality and moving speed of the vehicle. To be specific, the more closer the vehicle speed is to the average cluster speed, the longer this CH candidate would be stay in this cluster and the better it can serve as a cluster head; Similarly, the better the channel quality between CH and BS, the more reliable the trunk link transmission would be.

Note that the selected CH collects the status (position, velocity and heading direction) of vehicles, i.e., floating car data (FCD) and reports to the BSs. This kind of data is characterized as high frequency and small data size, which occupies cellular network resource frequently and impair other applications. Through clustering mechanism, FCD data is compressed and only transmits through the CH. However, this design increases the vulnerability of the system and poses a potential risk that the CH could be a single point of failure.

For this reason, we further propose a dual CH design in each cluster for SDN-enabled 5G-VANET to improve network robustness and guarantee the seamless communication during CH handover. In this dual CH scheme, a backup CH is also selected according to the CH selection criteria. The existing CH always sends a copy of FCD data to the back-up CH, as shown in Fig. 4.3. Once there is something wrong with the CH, such as an accident or unpredictable emergencies, backup CH would be well prepared and thus is able to take over the responsibility seamlessly. Moreover, the backup CH also works as a smooth transition during handover procedure to a new CH. This is to say, under the scenario that the existing CH leaves cluster normally, backup CH becomes CH immediately and a new back-up CH would be selected, as we can see in Fig. 4.3. The dual CH design is especially beneficial for 5G latency-stringent application with reduced communication interruption probability.

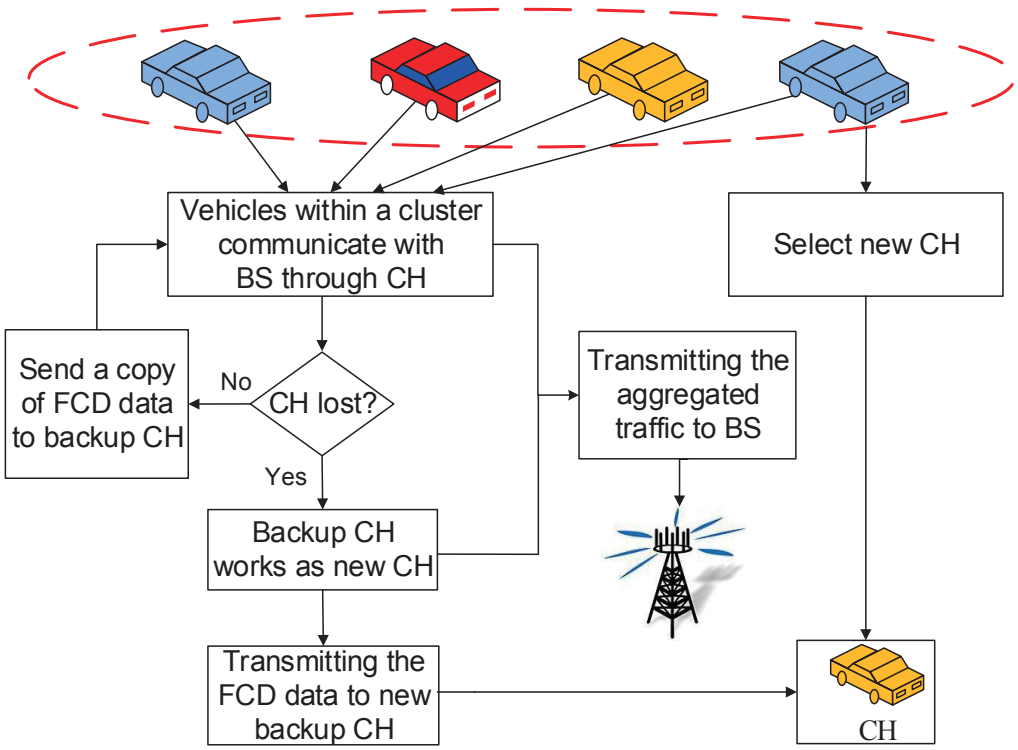

Figure 4.3: The dual cluster head selection scheme.

In order to show the performance of the aforementioned dual CH design, we derive the blocking probability of the CH transmission in an analytical way. It is assumed that arriving traffic from vehicles follows the Poisson distribution, while the service time is a exponential distribution with service rate μ equal to 1 per second. The resource requirement of the vehicles' transmission in different channel conditions are modeled as n independent one-dimensional Markovian models. Therefore, the total RB consumption can be modeled as the combine of n one-dimensional Markovian models [75]. If the RB consumption is larger than the free RB number in cell, the user requirement will be rejected and service is blocked.

Using the dimension reduction method provided in [76], we can derive the probability that m RBs are used by users with stationary distribution $q(n)$, which is a recursive formula:

$$q(n) = \sum_{s=1}^{S} \frac{\lambda_s \cdot \mu_s \cdot RB_s}{n} \cdot q(n - RB_s), \tag{4.4}$$
$$\sum_{s=1}^{N} q(n) = 1, n \geq 0, s = 0, 1, ..., S$$

S is the number of service types, i.e., the dimensions of the Markovian models. RB_s is the number of RB required by type s. The blocking probability P_b of type s is derived as:

$$P_b = \sum_{n=N-RB_s+1}^{N} q(n),\ n = 1, 2, ..., N \tag{4.5}$$

N is the number of RB of the cell. Formula 4.4 and 4.5 can be used in calculating the blocking probability of the vehicles to cellular traffic in terms of different traffic density.

4) Cluster maintenance and adaptation: Last but not the least, the clusters should be maintained and updated due to frequent road traffic changes in VANET. In the proposed SDN-enabled adaptive clustering scheme, the base station would only inform the corresponding CH if the new arriving vehicles will stay in the CH transmission area for a period larger than a threshold T_p. The predicted inhabitant time (PIT) is calculated using the angle of the new arriving vehicle to the center of the cluster and the speed of the arriving vehicle:

$$PIT = \frac{2 \cdot R_t(1 - \varepsilon) \cdot cos\theta_R}{V_i},\ 0^o < \theta_R < 90^o \tag{4.6}$$

θ_R is the angle of the new arriving vehicle to the center of the cluster, V_i is the speed of the arriving vehicle. In this way, the situation that fast moving vehicles are outside of the cluster coverage area before receiving information from the CH is avoided. Afterward, the CH would then be prepared for the new traffic and execute clustering algorithm only when needed.

On the other hand, if the aggregated amount of traffic exceeds the trunk-link capacity, the communication quality would deteriorate and outage probability will increase. Under this situation, some vehicles with high traffic requirement should be removed from the cluster to guarantee the communication Quality of Service (QoS). The cluster maintenance and adaptation

should be a monitored and on-going procedure regarding the communication quality index, e.g., outage probability.

4.4 Beamformed Adaptive Transmission Schemes in 5G-VANET

After the formation of vehicle clusters and the selection of CH, the CH would aggregate the traffic from other vehicles in the cluster and communicate with the cellular BS. As the volume of the aggregated traffic is much higher, provisioning of high capacity trunk link is critical in order to guarantee the communication performance of the clustered vehicles in the SDN-enabled 5G-VANET vehicle clustering design. Therefore in this section, beamformed adaptive transmission scheme of the trunk link between CH and BS will be elaborated in detail. Beamforming technique is used to provide directional coverage of the vehicle clusters with two selective coverage mode. Afterward, the adaptive trunk link transmission scheme is introduced, which consists of dynamic modulation and power control. When the trunk link traffic volume at the CH or latency requirement of the traffic is exceptionally high, cooperative communication of the mobile gateway candidates is also proposed in order to improve communication quality utilizing diversity gain and reduce the delay of traffic distribution through multi-user decoding.

4.4.1 Directional Coverage of Vehicle Clusters with Beamforming

After vehicle cluster formation, massive MIMO antennas are selected to form a highly directional beam to cover the given cluster or CH. Specifically, the macro BS estimates downlink channel information via uplink pilots under the assumption of channel reciprocity. In high mobility scenarios, use of channel reciprocity between uplink and downlink could introduce large error due to the reduced channel coherence time. In order to improve the system adaptivity, long-term channel prediction could be adopted to cope with the fast channel changes. According to these channel information, desired antennas would be selected to design beamforming gain vectors. Using linear precoders such as Minimum Mean Square Error (MMSE), Zero Forcing (ZF) and Maximum Ratio Transmission (MRT) precoding [77], the vehicles in different beamforming sub-bands would be orthogonal in space domain so that their interference is reduced dramatically.

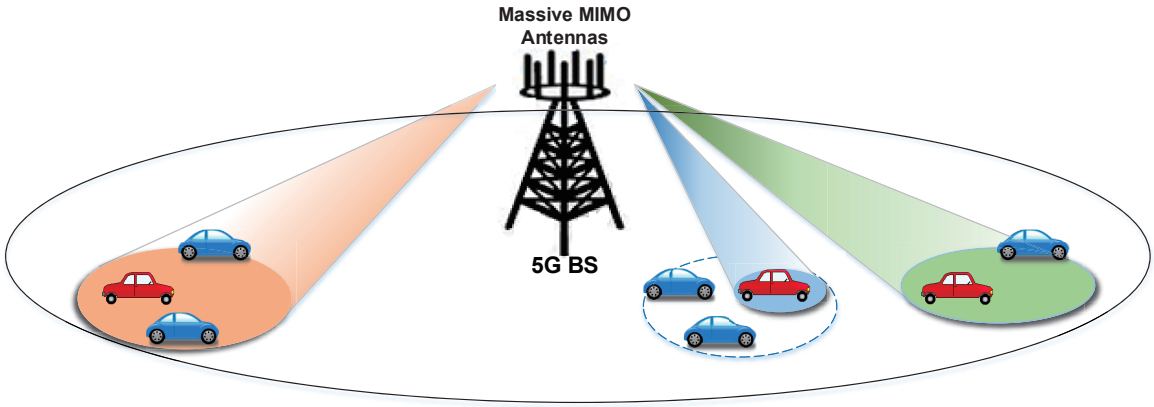

Figure 4.4: The beamforming design and directional coverage cover the vehicle clusters along a road crossing the cell.

The two beamforming solutions are shown in the right part of Fig. 4.4 further provide flexibility and adaptivity for the beamforming design of SDN-enabled 5G-VANET. In the first solution, the beam only covers the cluster head to optimize the beam width and reduce intra-beam interference. In the second technology, the beam covers the whole cluster. Although a wider beam in the latter scheme is less focused and achieves lower signal to interference and noise ratio (SINR) than a more focused beam, it has the advantage of better coverage and guarantees the seamless connection during CH failure. Therefore, due to the high mobility of vehicles, it is preferred that wider beam is utilized to enhance coverage in the proposed SDN-enabled 5G-VANET architecture. Furthermore, when there are multiple clusters co-exists and close to each other, the narrow beam will be applied to reduce interference and improve the trunk link throughput rate.

4.4.2 Adaptive Trunk Link Transmission for Aggregated traffic

As it is obvious that the amount of the V2I traffic would be changing at the different time, an adaptive transmission scheme for the aggregated traffic is designed in this section. The adaptive transmission scheme consists of adaptive modulation and coding (AMC) and power control. Compared with traditional AMC, we introduce a non-orthogonal multiplexed modulation scheme to cope with the varying channel condition of fast moving vehicles. In the proposed adaptive AMC, orthogonal and non-orthogonal modulation and coding scheme are selected adaptively according to channel quality in order to achieve high spectral efficiency.

Power control is also used to adapt power allocation to instantaneous channel variations so that a required SINR level can be guaranteed. As quality of service (QoS) is essential for real-time services, such as live video and interactive game, an optimal combination of AMC and power control techniques is used in the adaptive trunk link transmission scheme in order to achieve highest trunk link throughput rate, while maintaining the required QoS (measured by outage probability).

During the adaptive trunk link transmission, the CH uses traditional orthogonal modulation to communicate with the BS. If the trunk link traffic volume becomes higher than a threshold (e.g., when intra-cluster vehicles all request high data rate at the same time), non-orthogonal multiplexed modulation (NOMM) would be utilized to improve the trunk-link capacity and reduce the average transmit power. NOMM allows parallel data streams of one user to be modulated simultaneously and partially overlapped on a group of resource elements through sparse spreading code [78]. Compared with orthogonal modulation, NOMM is robust to the varying channel condition due to the fact that data streams belong to the same user suffered same channel variation. It can also improve the spectrum efficiency through overlapping on resource blocks. Therefore, in order to find a trade-off between throughput rate and receiver complexity, NOMM and orthogonal modulation should be selected adaptively according to varying traffic requirement.

To compare the throughput rate of different transmission schemes, we define the theoretical throughput rate as the successfully received bits, which can be approximated as:

$$Throu = bW(1 - \xi) \tag{4.7}$$

For traditional M-QAM modulation, $b = \log_{2^M}$ denotes the bits number of per symbol with M modulation order and ξ is the corresponding symbol error rate (SER). For NOMM modulation, $b = K \log_{2^M}$ denotes the bits number per codeword with M modulation order and K layers. ξ is the corresponding word error (WER) [78].

4.4.3 Cooperative Communication in 5G-VANET

A cooperation scheme of virtual MIMO based VANET is introduced in this section to further improve the communication quality and reduce the latency of traffic distribution phase when trunk link traffic amount is high. Next, the cooperative communication scheme will be explained in detail.

When the trunk-link traffic volume of a vehicle cluster head is higher than a threshold, cooperative communication is triggered and several mobile gateway candidates in that cluster, which have better channel quality, would be selected to share their antennas with CH as virtual antenna arrays and then communicate with the BS. The number of the vehicles participating in the cooperative communication is a trade-off between the performance and complexity. In the uplink transmission, the selected mobile gateway candidates serve as the sub-cluster head and collect traffic from the vehicles nearby. They transmit traffic to the BS simultaneously with the CH and thus improve the throughput with multiple trunk links; in downlink transmission, as shown in Fig. 4.2, the selected mobile gateway candidates will also listen to the BS and help the CH in traffic decoding, and thus reduce the latency of traffic distribution through multi-user cooperation. Therefore, the cooperative communication not only brings diversity gain but also potentially reduce latency in decoding.

4.5 Performance Evaluation

4.5.1 SDN Processing Latency

In order to evaluate the performance of the proposed SDN-enabled adaptive clustering algorithm compared with traditional methods, SDN controller processing ability is simulated in Matlab [79]. Without loss of generality, we assume that the Internet arriving data follows the Poisson distribution. In the Matlab simulation, SDN-enabled 5G-VANET predicts the vehicles moving direction and PIT before executing clustering, while traditional method periodically executes clustering. Despite the advantage of reduced computing cost compared with periodically clustering method, we hope to make sure that there is no extra latency introduced by SDN-enabled data traffic processing. Here we use two publicly available OpenFlow con-

trollers' processing data as a representative to show the performance of SDN controller[69], i.e., NOX-MT and Beacon. NOX-MT is a multi-threaded successor of NOX, while Beacon is a Java controller built by David Erickson at Stanford [21].

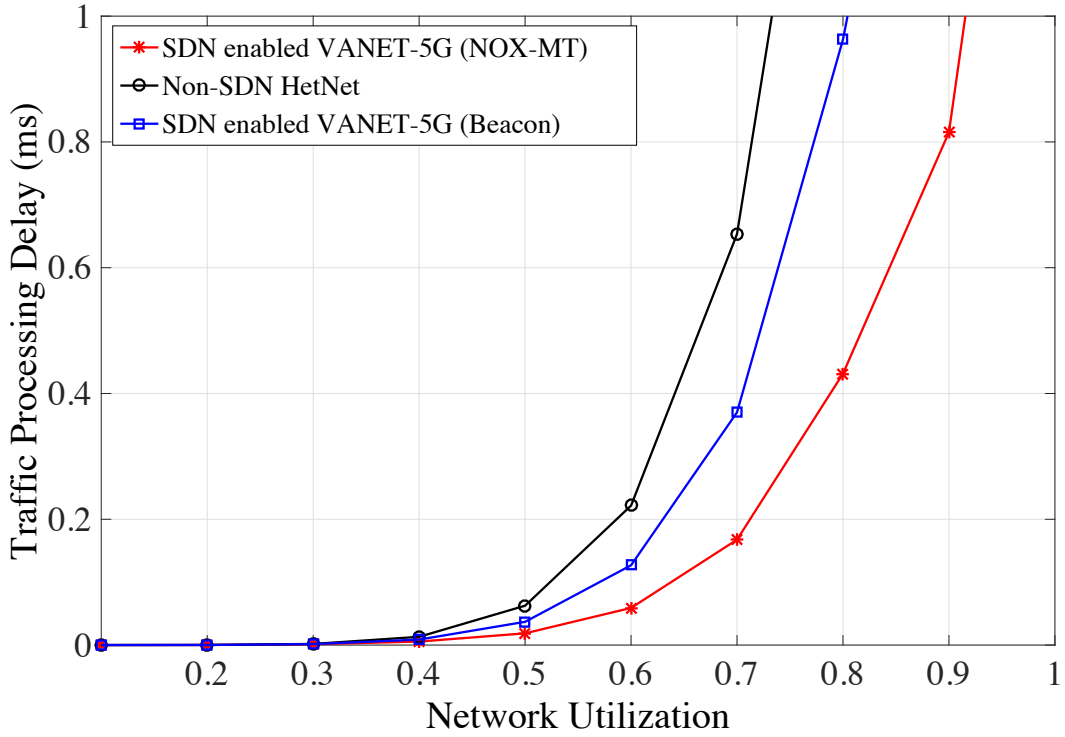

Figure 4.5: Simulation results of SDN-enabled 5G-VANET delay compared with Non-SDN networks in terms of network utilization.

Fig. 4.5 shows the comparison of processing delay versus network utilization rates. Here network utilization is defined as the ratio of total data arrival rate and the controller processing rate. Network utilization rate is used as a reflection of the various load situation of the network in order to provide more accurate latency performance. It can be seen from Fig. 4.5 that when the network load is fairly low, processing delay is not a problem for both SDN and non-SDN networks. With more data arrivals and increased network load, SDN-enabled 5G-VANET still keeps the latency under $1ms$ most of the time, which meets the 5G latency requirement. NOX-MT and Beacon-enabled solutions perform 30% and 14.29% better than traditional method in latency reduction with the commonly used deployment of an eight-core machine, 2GHz CPUs and 32 switches in [69]. It is obvious that SDN-enabled 5G-VANET has better performance in meeting the critical latency requirement in 5G, while maintaining the SDN flexibility, pro-

grammability for 5G networks.

4.5.2 SDN-enabled Adaptive Clustering and Dual Cluster Head Selection

To evaluate the performance of SDN-enabled 5G-VANET applications, MATLAB simulations have been conducted regarding the SDN-enabled adaptive clustering scheme and adaptive transmission scheme of aggregated traffic. We present the BER-SNR performance with QPSK modulation, and the channel coefficients matrix is generated by Jake's model. Note that using the principle of Monte Carlo, all simulations are carried with a 95% Confidence Interval (CI) [81]. SDN-enabled adaptive clustering scheme is compared with existing mechanisms and three scenarios have been considered in the simulations: the proposed scheme which uses SNR combined with average speed to select CH; traditional method which chooses center vehicle as CH, and the scenario that there is no clustering mechanism (vehicles communicate with BSs through their connection).

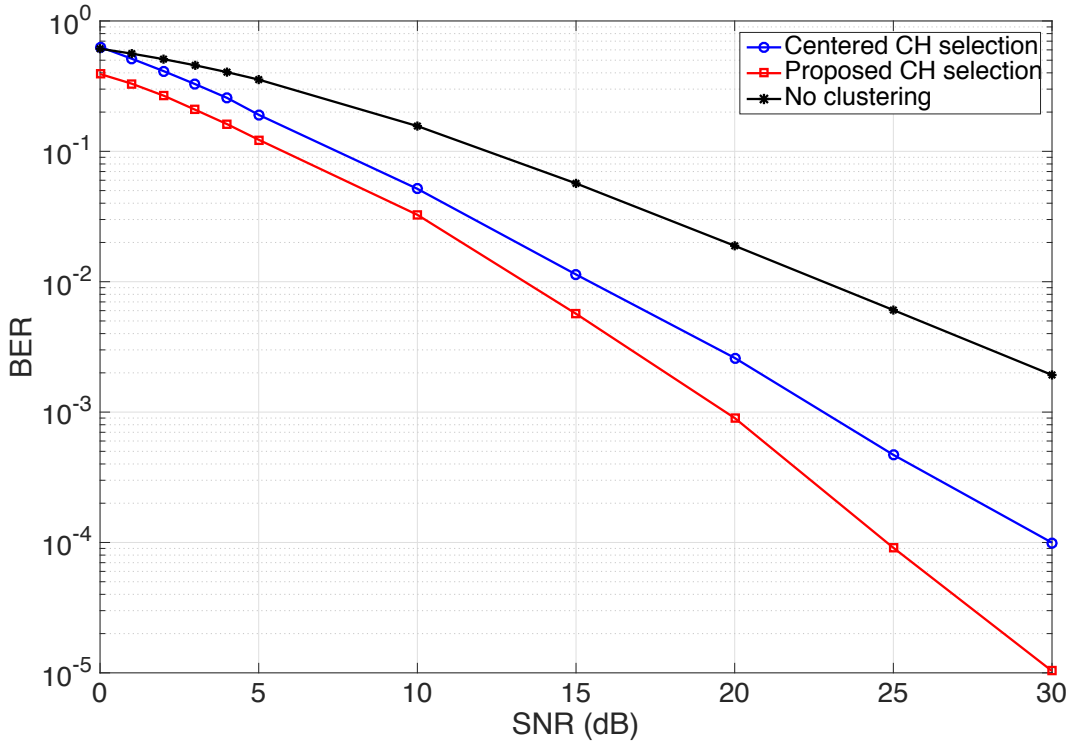

Figure 4.6: Simulation results of BER vs SNR in terms of three different vehicle clustering and CH selection methods.

Fig. 4.6 illustrates the simulation results of bit error rate (BER) vs. SNR regarding the three

different clustering and CH selection scenarios. It can be seen that clustering provides better communication quality for vehicles, which is reasonable because vehicle clustering schemes use IEEE 802.11p network to offload the burden of the cellular network, and thus guarantee the mobile radio link connection quality. It is also clear that the proposed CH selection scheme (i.e., the red line in the figure) has the lowest trunk link BER, especially in higher SNR situation. This is because the proposed CH selection scheme takes both SNR and average speed into consideration when choosing a CH, and thus end up with an optimum CH that has better radio link quality and serves longer in the cluster (less CH handover).

Next, the blocking probability of FCD data transmission [73] is simulated and compared between dual CH design, single CH scheme, and no clustering schemes to show the performance. The term "blocking probability" here is defined as the probability that the RB resource required by the users exceeds the RB resource number in a cellular cell. In simulation setup, we assume that in dual CH scheme, the CH relay communication is seamless due to the design of a backup CH (elaborated in Fig. 4.3), while in previous single CH scheme, vehicles communicate with BSs directly during new CH selection period. In no clustering scenario, vehicles always transmitting FCD data with BSs directly.

In Fig. 4.7, we can see that generally, the probability of blocking increases when there is more arriving traffic. This is to say, service disruption would happen when the network gets busier. For the case of clustering, existing CH is also more likely to have an accident or get lost on busy roads in reality. In all the three scenarios, more arriving traffic means more burden on the cellular resource and thus higher blocking rate. It is also clear from Fig. 4.7 that clustering schemes have lower blocking probability than no clustering scheme, where vehicles set up multiple communication links with cellular BSs directly. Moreover, compared with single CH scheme, dual CH design reduce blocking probability to some extent in the case that existing CH is lost suddenly.

4.5.3 SDN-enabled Beamformed Adaptive Transmission Scheme

In the simulation of adaptive transmission scheme, QPSK and NOMM are simulated as two modulation scheme to show the transmission performance. It is assumed that there are 6 ve-

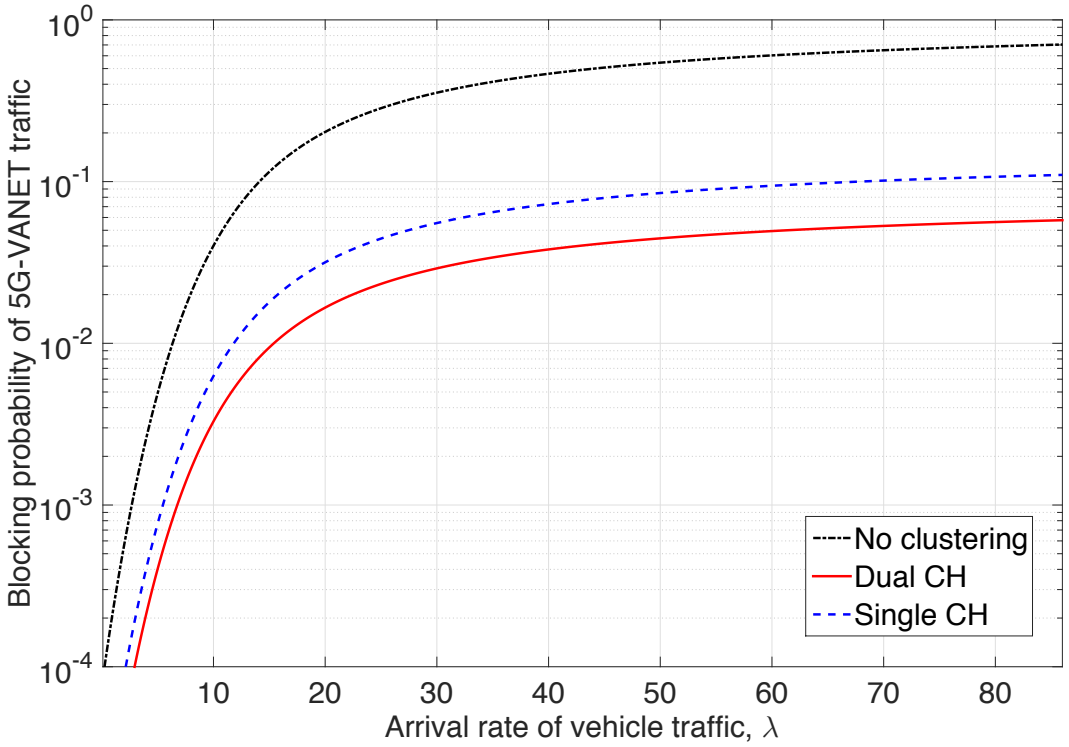

Figure 4.7: Blocking probability of 5G-VANET link vs. arrival rate of vehicle traffic.

hicles randomly distributed within a cluster and the data packet size from each vehicle is 512 Bytes [80]. Each vehicle generates 10 packets per second. In NOMM scheme, QPSK modulation combined with sparse spreading were used for each layer and ML detection was applied to the receivers. Link level simulation is implemented including channel coding, modulation, and demodulation. Monte Carlo simulation is also given in order to verify the theoretical analysis.

According to [82], there are two important factors for NOMM signature design: the minimum euclidean distance and the girth (the minimum cycle length of factor graph). The detection performance becomes better with the increase of the minimum euclidean distance and the decrease of the minimum cycle length. The principle of optimal signature design for a given factor graph is to find a trade-off between the minimum Euclidean distance and the minimum cycle length. Using this principle, some optimal signature matrix examples are designed for the simulation, as given below:

Example 1 (An Optimal 8-Layers, 4 Resources NOMM Code):

$$S_{4,8}^{opt} = \begin{bmatrix} 1 & 0 & e^{i\theta_2} & 0 & e^{i\theta_4} & 0 & 0 & 0 \\ 0 & e^{i\theta_1} & 0 & e^{i\theta_3} & 0 & e^{i\theta_5} & 0 & 0 \\ 0 & 0 & e^{i\theta_2} & 0 & 0 & e^{i\theta_{3,6}} & e^{i\theta_6} & 0 \\ 0 & 0 & 0 & e^{i\theta_3} & e^{i\theta_4} & 0 & 0 & e^{i\theta_7} \end{bmatrix} \tag{4.8}$$

Where $(\theta_1, \cdots, \theta_7) = (0.2618\pi, 0.1435\pi, 0.1279\pi, 0.2297\pi, 0.3505\pi, 0.3935\pi, 0.361\pi)$ and $\theta_{3,6} = 0.2269\pi$. Its the minimum code distance is 0.83 and load ratio is 2.

Example 2 (An Optimal 6-Layers, 4 Resources NOMM Code):

$$S_{4,6}^{opt} = \begin{bmatrix} 1 & e^{i\theta_2} & e^{i\theta_3} & 0 & 0 & 0 \\ 1 & 0 & 0 & e^{i\theta_4} & e^{i\theta_5} & 0 \\ 0 & e^{i\theta_2} & 0 & e^{i\theta_{3,4}} & & e^{i\theta_6} \\ 0 & 0 & e^{i\theta_3} & 0 & e^{i\theta_{4,5}} & e^{i\theta_{4,6}} \end{bmatrix} \tag{4.9}$$

Where $(\theta_2, \cdots, \theta_6) = (0.1431\pi, 0.2021\pi, 0.3127\pi, 0.3765\pi,\ 0.2667\pi), \theta_{3,4} = 0.5736\pi, \theta_{4,5} = 0.3935\pi$ and $\theta_{4,6} = 0.3078\pi$. Its minimum code distance is 1.1658 and load ratio is 1.5.

Example 3 (An Optimal 6-Layers, 4 Resources NOMM Code):

$$S_{4,6}^{opt} = \begin{bmatrix} 1 & 0 & e^{\frac{i\pi}{6}} & 0 & 0 & e^{\frac{i\pi}{6}} \\ 0 & 1 & 0 & e^{\frac{i\pi}{6}} & e^{\frac{i\pi}{3}} & 0 \\ 0 & 0 & e^{\frac{i\pi}{6}} & 0 & e^{\frac{i\pi}{3}} & 0 \\ 0 & 0 & 0 & e^{\frac{i\pi}{6}} & 0 & -1 \end{bmatrix} \tag{4.10}$$

The minimum code distance is 1.2679 and load ratio is 1.5. Compared with example 2, its minimum code distance is larger and so the performance of ML detection is better.

Example 4(An Optimal 8-Layers, 6 Resources NOMM Code:

$$S_{6,8}^{opt} = \begin{bmatrix} 1 & 0 & e^{\frac{i\pi}{6}} & 0 & 0 & 0 & 0 & e^{\frac{i\pi}{6}} \\ 0 & e^{\frac{i\pi}{3}} & 0 & e^{\frac{i\pi}{6}} & 0 & 0 & e^{\frac{i\pi}{6}} & 0 \\ 0 & 0 & e^{\frac{i\pi}{6}} & 0 & e^{\frac{i\pi}{3}} & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{\frac{i\pi}{6}} & 0 & e^{\frac{i\pi}{3}} & 0 & 0 \\ 0 & 0 & 0 & 0 & e^{\frac{i\pi}{3}} & 0 & e^{\frac{i\pi}{6}} & 0 \\ 0 & 0 & 0 & 0 & 0 & e^{\frac{i\pi}{3}} & 0 & e^{\frac{i\pi}{6}} \end{bmatrix} \tag{4.11}$$

Its minimum code distance is 1.412 and load ratio is 1.33.

The load ratio of Example 1 to 4 is 2, 1.5, 1.5, and 1.33, while the minimum code distances are 0.83, 1.166, 1.27 and 1.41, respectively [83].

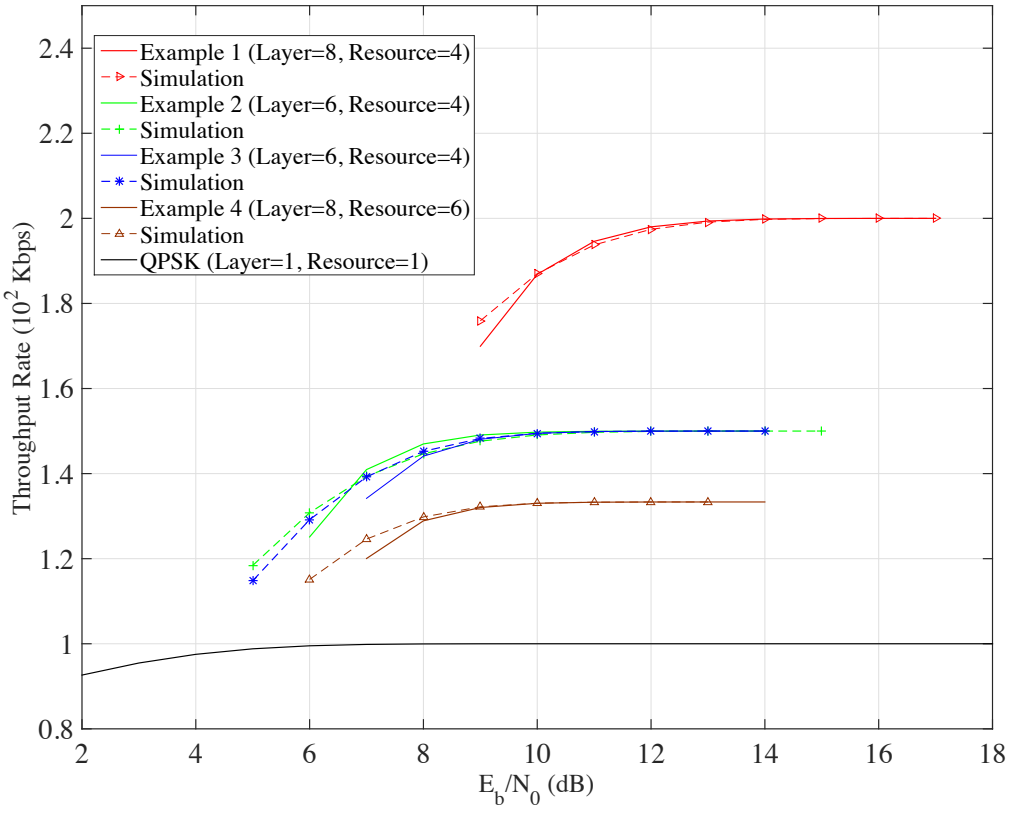

Figure 4.8: Throughput rate comparison of two trunk link modulation schemes: NOMM modulator and QPSK modulation.

Fig. 4.8 illustrates the throughput rate (dashed lines) and the union bounds (solid lines) [82] of NOMM codes given in Example 1, 2, 3, 4 [83], and single user QPSK code. From the simulation results, we can see that: 1) all the simulations coincide well with their union bound most of the time, except a little mismatch at the low $\frac{E_b}{N_0}$ due to noise; 2) For NOMM examples with different bits number per symbol, the throughput rate becomes higher with increased bits number per symbol. For example, code obtained in Example 1 with the optimal signature has the best throughput rate since it has the maximum bits number per symbol; 3) the throughput rate of NOMM is better than that of QPSK, due to the overlay transmission. That is to say, NOMM provides higher throughput rate at the cost of complexity. Therefore, adaptive transmission of the aggregated traffic should work in a way that QPSK or NOMM is selected dynamically according to traffic requirement.

4.6 Chapter Summary

With the anticipated arrival of self-driving vehicles and dramatic growth of in-vehicle mobile data traffic, supporting of dynamic vehicle communications in 5G HetNets are expected to be extremely challenging, due to fast varying network topology and high complexity of the heterogeneous infrastructure. In this chapter, we propose to integrate SDN into 5G-VANET and thus provide a programmable platform in addressing the challenges of dynamic vehicle communications. Through the proposed SDN-enabled adaptive vehicle clustering and dual cluster head scheme, signaling overhead of VANET is significantly reduced along with improved communication quality. The proposed cluster head selection also guarantees the seamless access to the operators' services for the cluster users. To accommodate the varying traffic over the trunk link and reduce the latency during traffic distribution, adaptive trunk link transmission scheme and cooperative communication of mobile gateway candidates were proposed for the aggregated V2I traffic transmission in this integrated network. Simulation results show that SDN coordinated vehicle clustering and beamformed transmission are suitable to support fast varying traffic conditions with extremely extended dynamic range.

Chapter 5

SDN-enabled Orchestrated Spectrum Sharing in 5G HetNets

5.1 Introduction

With the expected growing spectrum scarcity due to data-consuming smartphones, laptops, and tablets, techniques for improving spectrum efficiency have received tremendous interest during the past decades and are considered to be critical performance indicator of future wireless networks. According to Cisco's networking visual index report [6], the mobile data traffic is expected to grow at a compound annual rate of 61% by 2018, an 11-fold increase over 2013, where mobile video traffic accounted for 55% of overall traffic. Moreover, besides the fact that the number of traffic-intensive mobile devices are increasing rapidly, the induced user data traffic has more stringent quality of service (QoS) requirements due to the diversity of applications [84]. Given the dramatic growth of mobile traffic, very limited spectral resources and increasing QoS provisioning challenges, efficient spectrum management, and allocation mechanisms becomes essential for future wireless networks to fully utilize all available spectrum resources while supporting the increased data rate.

There have been lots of related studies in addressing the spectrum shortage problem in the last twenty years. One direction of such efforts is to allocate new spectrum resources, i.e., exploring unused spectrum particularly in the millimeter wave (mmW) band of $30 \sim 300GHz$ [9]. However, it is tough to provide broad coverage in mmW band due to the poor signal

propagation characteristics at extremely high frequencies. Another direction is to improve the spectrum utilization rate through spectrum sharing among co-existing wireless networks using dynamic spectrum sharing or cognitive radio (CR) technology. As some frequency bands in the existing spectrum allocation are underutilized, for example, UHF/VHF bands (which is around 470-698 MHZ) reserved for broadcasting, it is beneficial to explore the spectrum white spaces for mobile data communications due to their desirable propagation characteristics to support coverage and mobility [84]. These TV white space could be aggregated properly to guarantee the required bandwidth. As a result, CR was first proposed about twenty years ago to enable secondary access to the unused spectrum while reducing the possible interference among the multitude of wireless systems through spectrum sensing. However, due to the limited sensing capability of devices and lack of timely information exchange between coexisting devices and networks, it is extremely difficult to avoid misdetection and performance degradation of the primary users [10]. Due to these risks, operators of primary networks are reluctant to share their spectrum resources with secondary networks. The low industry interest on standardization or production of CR technology in the industry also demonstrates these practical concern.

In general, conventional spectrum sharing schemes, particularly those with over-reliance on spectrum sensing, mostly are based on limited input and link level decision-making by individual devices without coordination across co-existing networks or an efficient admission/eviction mechanism for secondary users. These challenges are further compounded by the fact that existing broadcasting and broadband wireless networks have a rigid architecture relying on vendor-specific configuration and air interfaces. There is hardly or no effective, timely information sharing among the independent, heterogeneous networks for the real-time spectrum access. Thanks to the ongoing development of the next-generation ATSC 3.0 DTV broadcast system with return channel [85] and the effort of opening more under-utilized spectrum for sharing, the convergence of broadcasting and broadband wireless systems becomes possible and beneficial, as the return link enables reliable spectrum occupancy information collection through real-time updates from incumbents.

In this chapter, we adopt an orchestrated spectrum sharing approach that integrates the distributed located users, base stations (BS), incumbent stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with the real-time exchange through

Internet. To effectively protect incumbent users and efficiently share the pooled spectrum resources, real-time spectrum information regarding the 3D interference map will be considered for new spectrum access with SDN' global control and coordination capability among HetNets. With spectrum occupancy inputs from incumbent systems and secondary user sensing reports, a 3D interference map is generated to guide the spectrum sharing procedure. Secondary users then access the shared spectrum only when the interference level is within threshold to guarantee overall network performance. Protected spectrum sharing improves spectrum utilization by allowing multiple users, both government, and commercial, to dynamically access the spectrum resource in a controlled way. We also propose a simplified layered SDN-enabled HetNet, where SDN only define high-level rules and leave the decision making to base stations (BSs)/ incumbent stations.

SDN is a new programmable network structure in light of the recent advancement of computing and network virtualization technology. It will be able to integrate the control entities of underlying infrastructures of broadcasting and wireless systems to a controller in control layer to realize global spectrum management. This way, the rigid network architecture can be innovated into a programmable common interface for spectrum information exchange and configuration. Instead of indirect control of the heterogeneous networks, the controller is now aware of the real-time spectrum utilization, and the unused spectrum resources are reported and pooled for sharing.

In this way, we achieve a three-fold advantages:

- BS/ incumbent stations make best decision with most of the knowledge. In wireless environment that is experiencing attenuation and fading, BSs and incumbent stations can make the best decision because they could collect real-time information from devices;

- Reduce latency and robust to failure. As SDN only define the high-level spectrum sharing rules, the processing time and decision updating time of traditional SDN network are dramatically reduced. Even when there is a failure of the SDN controller, the BSs and incumbent stations still can work as usual;

- More efficient spectrum utilization and improved user experience. Device level decision making introduces interference and security threats. With orchestrated spectrum shar-

ing using 3D interference map, incumbents are protected, and secondary users achieve reliable opportunistic spectrum access.

The remainder of this chapter is organized as follows: Section 5.2 and 5.3 presents the SDN-enabled orchestrated spectrum sharing scheme using 3D interference map, including SDN system model and spectrum sharing scheme design. Further, performance evaluation is presented in Section 5.4. Finally, the conclusion is drawn in Section 5.5.

5.2 System Model

5.2.1 Layered System Model based on SDN

The future wireless communication lies in reliable sharing of spectrum across time, space and other dimensions [86]. The spectrum resource is also expected to be described as a distribution with different interference level instead of simple "Yes" or "No" decision. In doing so, high-level spectrum sharing rules are needed in order to effectively protect the incumbent system who own the spectrum and guide the access from secondary users. Therefore in this article, we introduce Software-defined Networking technology as an enabling platform to apply intelligence and programmability into the spectrum sharing procedure. SDN is a layered structure, in which the control entities of underlying infrastructures are taken out to a controller in control layer [21]. Therefore, software applications can be written upon the controller to realize functions, like routing, resource allocation, and spectrum management. Fig. 5.1 below shows the system model of SDN-enabled orchestrated spectrum sharing.

In Fig. 5.1, a cluster of BSs/ TV stations are controlled by a local SDN controller through high-capacity fiber optic links and utilizes the OpenFlow protocol for controlling the data plane. Additionally, the Simple Network Management Protocol (SNMP) plug-in enables the controller and the relevant applications to monitor and manage the data plane devices [21]. The SDN controller is in charge of the policies related to various access network functionalities, including resource allocation, spectrum management, security provisioning [72], and so on, and these features form the control plane of the radio access network. The BSs constitute the data plane of the network and implement the controller-defined policies.

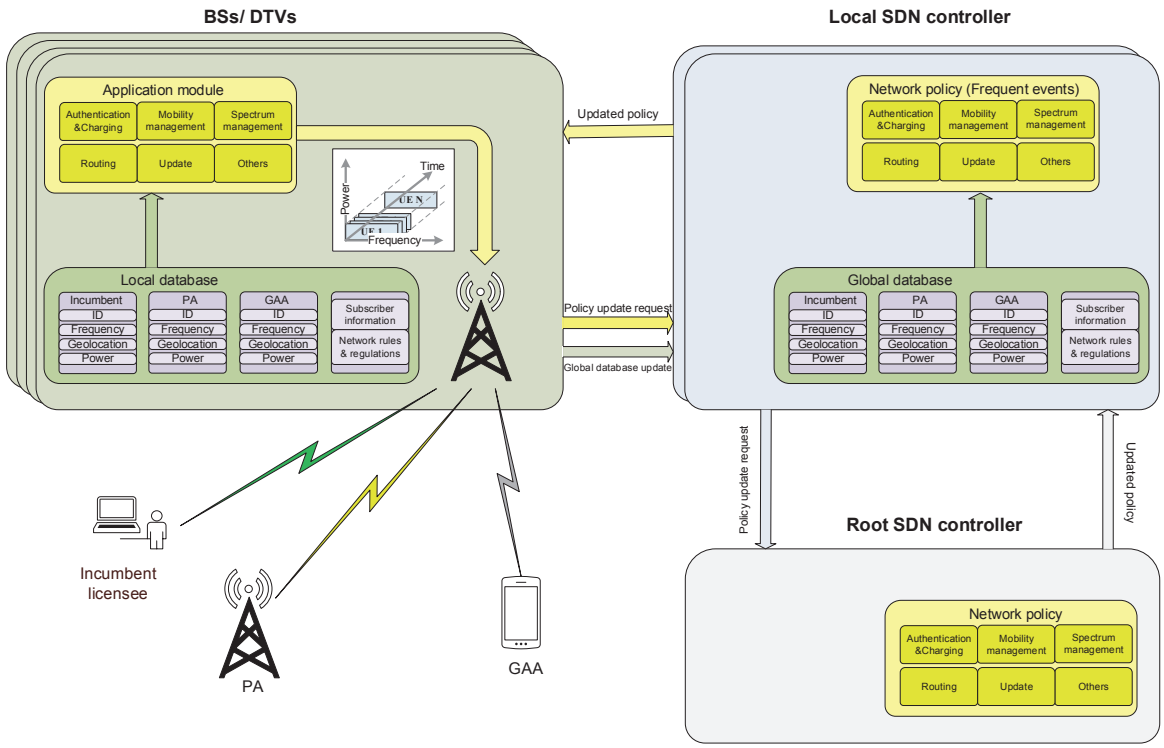

Figure 5.1: System model of SDN-enabled spectrum sharing.

There are three tiers of users in the spectrum sharing system with decreasing priority, namely, incumbent users, Priority Access (PA) users, and General Authorized Access (GAA) users:

Incumbent users Incumbents are those authorized federal, Fixed Satellite Service or DTV users currently operating in the frequency band. These users should be protected from harmful interference from the secondary access of PA and GAA users.

Priority Access users The PA tier users are those who were assigned the spectrum (e.g., 10MHz channel) for a limited time and geographical area after bidding.

General Authorized Access users The GAA tier is the part of users that are able to access the spectrum opportunistically if the spectrum is not occupied by any higher tier users.

SDN centralized control capability facilitates the HetNets management based on distributed input from user sensing reports and spectrum occupancy database. Due to its open interface and ease of reconfiguration, SDN provides network operators a programmable platform for global control over the entire HetNets instead of the vendor-specific configurations. Consider-

ing the scalability and latency issues caused by SDN centralized processing, layered-SDN is proposed in the system model, where the network management tasks are distributed between root SDN controller, local SDN controller and BSs/ TV stations [10]. To this end, the BSs/ TV stations carry out all the localized decision making according to the high-level policy defined by the local SDN controller. Local SDN controller managed the local spectrum sharing procedure, while the root SDN controller handles the overall spectrum utilization, such as the pool of spectrum resources that could be shared, protection zone settings for incumbent users, or updates spectrum bidding results with the local SDN controller. In this way, SDN controllers only manage the network rules and high-level policies, and BSs or incumbent systems such as TV stations takes care of local decision making such as handovers as they are closer to users and have the most up-to-date information under rapid varying channel conditions.

5.2.2 Shared Spectrum Access Model

Fig. 5.2 describes the procedure of shared spectrum access. The orchestrated spectrum sharing system includes incumbents, spectrum database, and local SDN controller. The incumbents update the real-time spectrum availability to the spectrum database, for example, next-generation ATSC 3.0 DTV broadcast system can use the return channel for usage feedbacks [85]. The BSs send spectrum requests to the local SDN controller if they need a short-term license in a geographical area (as PA tier access), and GAA users are also allowed to sense the spectrum and report their findings to their nearest BSs. The local SDN controller will then interrogate the spectrum database and check the 3D interference map to allocate the best spectrum back to the BSs/ GAA users.

After signing the agreement of spectrum sharing, the unused spectrum resources of incumbents will be added to pools for sharing. Institutions or operators could bid for fixed spectrum availability for a period or restricted location as PA users, while GAA users can have opportunistic access considering the interference level using 3D interference map. For example, the special events in certain localized areas (lasts for days or months) that do not need long-term spectrum license get the convenience of spectrum bidding and the unused portion of the licensed spectrum from incumbents such as DTV systems could be shared under coordination.

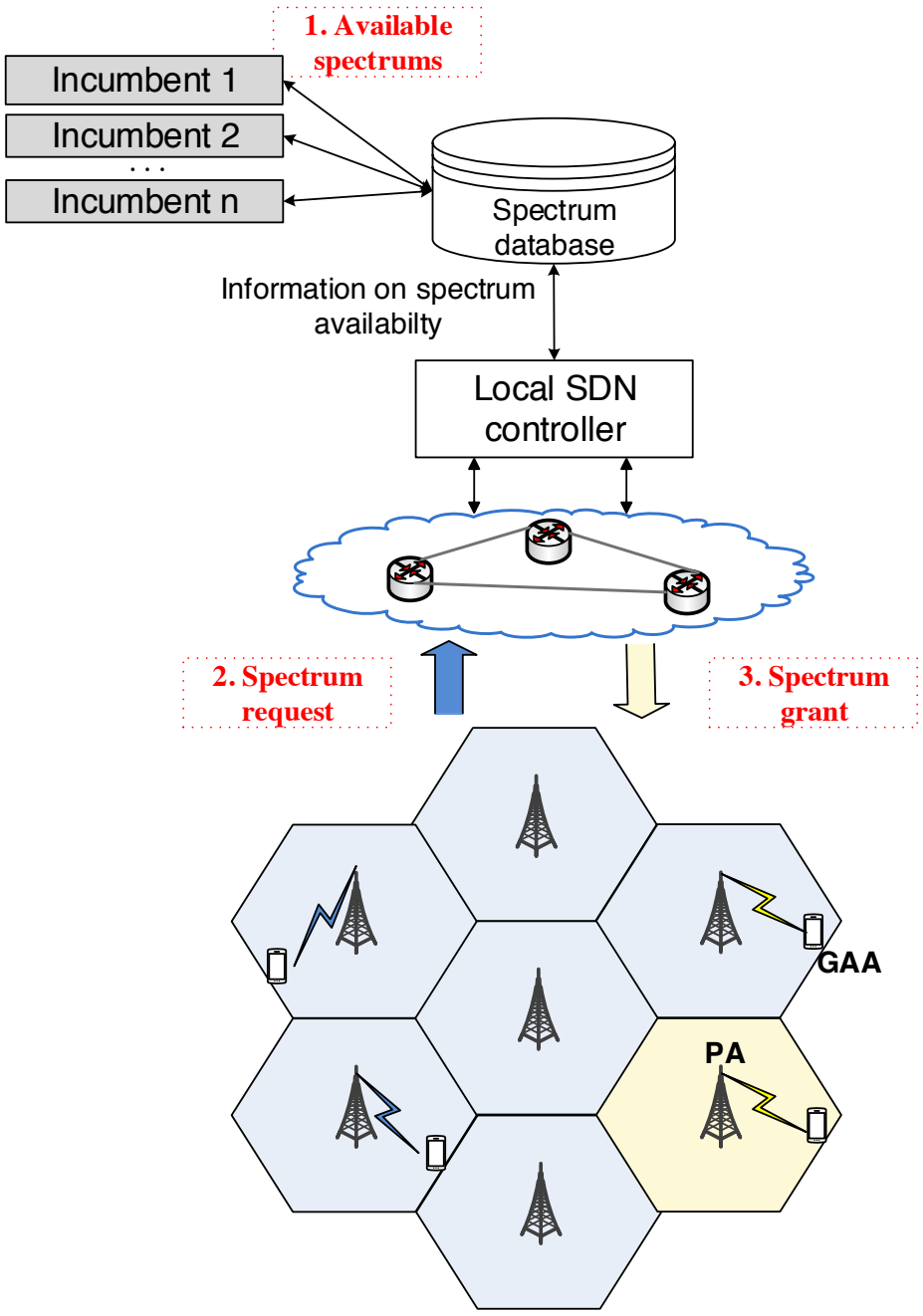

Figure 5.2: The procedure of shared spectrum access with the help of local SDN controller and Spectrum database.

The shared spectrum could also work as the bandwidth expansion for licensed band [86]. During the rush hours at a particular area, the BS applies for a temporary local authorization to use the shared spectrum. In this way, the orchestrated spectrum sharing offload the high traffic and release cellular burden with the flexible and on-demand spectrum sharing service.

5.3 SDN-enabled Orchestrated Spectrum Sharing using 3D Interference Map

5.3.1 3D Interference Map

There have been lots of related works regarding spectrum sharing recent years. Authors in [87] proposed spectrum sharing using licensed shared access (LSA) as a complementary approach to traditional exclusive usage. The overall workflow is studied, including LSA preparation, licensing, deployment, and release. In [10], a centralized spectrum sharing management platform based on SDN is introduced with policy define procedure and processing load distribution. However, a spectrum availability indicator is still to be developed to guide the spectrum sharing procedure under the pre-defined management platform.

In this section, orchestrated spectrum sharing under the SDN framework is realized using 3D interference map, as it permits dynamic spectrum allocation for both the incumbent and PA, GAA users without harmful interference. Note that here 3D means time, frequency and space. It is believed that the quality of spectrum resource is not simply decided by the channel quality or interference level measured just by the BSs or UEs. It is also influenced by stability, availability over time, channel condition over certain space, communication quality and the impact of co-existing users or networks. Therefore in this section, a 3D soft interference map is proposed to better describe and measure the licensed and unlicensed spectrum resources.

Time domain The consideration of time stability of a spectrum is essential in spectrum availability evaluation due to the dynamic nature of wireless communications and the ongoing evolution of wireless technologies, applications, and regulations in the sharing of licensed bands. For example, the prediction and awareness of the ON/OFF pattern of a candidate spectrum are critical to avoid channel changing and user service disruption during communication. Specifically, if a spectrum is owned by a higher priority system, it would be turned OFF whenever required by high priority system and the QoS of secondary systems is thus impaired. If a spectrum is not stable or heavily used for some time duration according to historical channel utilization data, it is better to be marked as low time stability during that period to avoid service disruption during spectrum sharing. Under this scenario, a trade-off should be considered in

choosing the candidate spectrum for sharing with best SINR, compared with a second choice which would be stable all the time during secondary transmission.

Channel condition over space Due to the varying propagation distance, the activity of interfering signals and impact of the surrounding environment, the channel quality measurement by only the two users/systems (i.e., BS or UE) is limited as interference from co-channel transmitters could be very different during user mobility. Even if the interference might be below a threshold at the time of measurement, it cannot guarantee the interference generated during the whole communication session. Therefore, the spatial distribution of candidate channels is taken into consideration as part of the 3D soft interference map calculation. Specifically, a buffer zone over space would be set up to give a reliable protection of the incumbents during user mobility.

Potential impact to co-existing users/network and communication quality Traditional spectrum sharing decisions were made in link level without considering the impact of potential spectrum sharing to co-existing users and networks. In the 3D interference map generation, the mutual influence among the homogeneous/heterogeneous networks would be modeled to ensure that the potential spectrum sharing don't impair the overall network level performance and decrease the cross network impact. Our objective is to obtain better spectrum efficiency through spectrum sharing, but without significant influence on co-existing homogeneous/heterogeneous networks.

The system model mentioned above in the last subsection could be taken as an example to better illustrate the three dimensions. In the time domain, PA or GAA users can have exclusive spectrum rights of use when the spectrum is not used by the incumbent. Incumbent and PA, GAA users may share the same spectrum in the same location during different time periods, thus there is no interference issue. In frequency and space domain, it is obvious that if incumbents and other tiers of users use the same spectrum at the same time in nearby locations, interference avoidance must be considered. The interference map is thus used as a guidance and protect the incumbent from harmful interference.

5.3.2 3D Interference Map based Orchestrated Spectrum Sharing

Fig. 5.3 shows the calculated interference map used during spectrum sharing procedure. As the channel condition varies over space at distributed locations, interference from co-channel transmitters could also be very different during user mobility due to the varying propagation distance, the activity of interfering signals and impact of the surrounding environment. Therefore, it is believed that spectrum could be shared for shorter time scale instead of simple binary decision for the whole communication session. For example, as shown in Fig. 5.3, user A is moving across an area that has spectrum sharing possibility and request for spectrum access. Assume that the channel model $\mathbf{h}_i \in \mathbb{C}^{1\times N}$ (N is the number of new access in total) between the BS that shares the spectrum and the incumbent user m is given by [88]

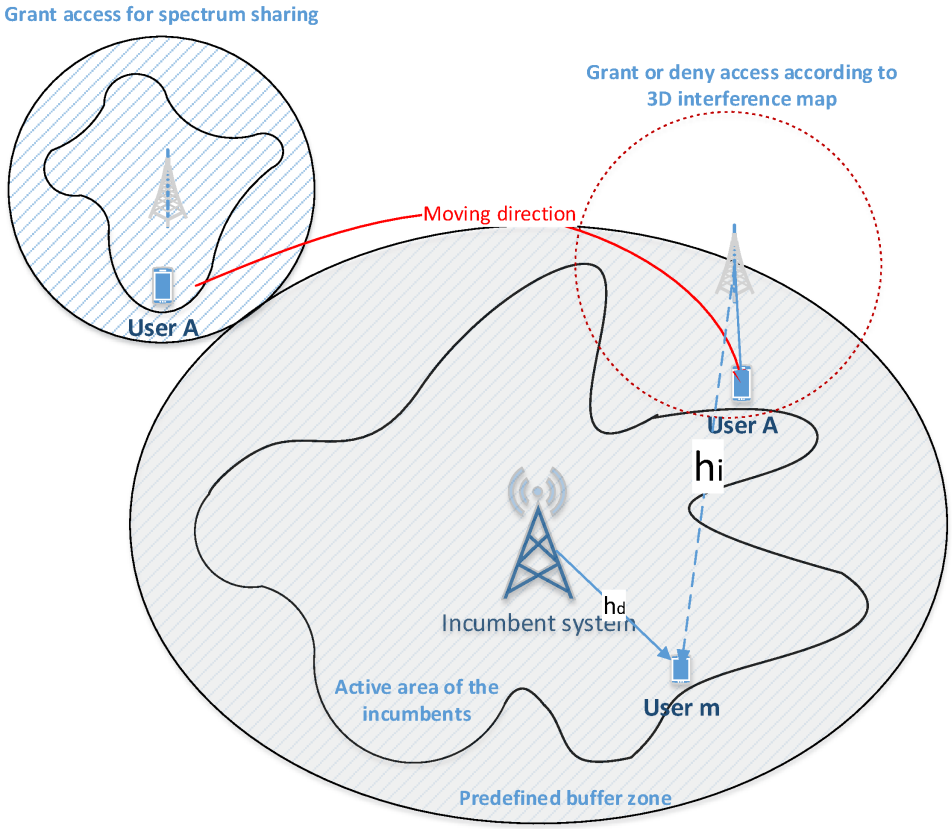

Figure 5.3: The interference map and buffer zone for the protection of existing users during spectrum sharing.

$$\mathbf{h}_i = \tilde{h}_i \sqrt{\mathbf{R_i} \cdot G \cdot A_i(\varphi, \theta) \cdot PL(d_i) \cdot Z_i \cdot \alpha_i} \tag{5.1}$$

where $\tilde{h}_i \sim \mathcal{CN}(0, \mathbf{I}_N)$ are independent fast fading channel vectors and the other part of the equation represents propagation gain consisting of the positive semidefinite spatial correlation matrix R_i, BS antenna gain G, the antenna radiation pattern in linear scale $A_i(\varphi, \theta)$ [88], path loss PL_{d_i}, the log-normal shadowing Z_i and the small-scale fading α_i. To be specific, R_i corresponds to ith path between interfering BS and incumbent m with a specific channel gain. R_i can be characterized for the commonly adopted uniform linear array (ULA) models according to IEEE 802.11p standard [89]:

$$\mathbf{R}_i = a(\theta_i)a(\theta_i)^H \tag{5.2}$$

Assume that there are K number of antennas at the BS that is communicating with the new access UE using shared spectrum from the incumbent m, θ_i is the angle of arrival (AoA) for the ith interfering path, then

$$a(\theta) = \frac{1}{\sqrt{K}}[1, e^{j\frac{2\pi}{\lambda}s \cdot sin(\theta)}, ..., e^{j(K-1)\frac{2\pi}{\lambda}s \cdot sin(\theta)}]^T \tag{5.3}$$

where s is the antenna spacing [89].

Assume that γ is the path loss exponent, we fit the path loss to a log-distance path loss model, as in [90]:

$$PL(d_i) = PL(d_0) + 10\gamma log_(d_i/d_0) + X_\sigma \tag{5.4}$$

where $\gamma = 2$ for free apace, and it is generally higher for wireless channels. X_σ is zero mean log-normally distributed random variable related with shadow fading. The term $PL(d_0)$ gives PL at a known reference distance d_0 which is the far field of the transmission antenna (typically 1 km for large urban mobile systems, 100m for microcells, and 1 m for indoor systems), which is $PL(d_0) = -G_t - G_r + 20log((4\pi f d_0)/c)$. Hence we obtain the path loss model as:

$$PL(d_i) = -G_t - G_r + 20log((4\pi f d_0)/c) + 10\gamma log(d_i/d_0) + X_\sigma \tag{5.5}$$

where c is the light speed, G_t and G_r is the antenna gain of the transmitter and receiver, and d_i is the distance between the nearby AP that shares the spectrum and the incumbent m.

Combining Eq. (5.2) and Eq. (5.5) into Eq. (5.1), the interference that user A might introduce at the incumbent user m can be derived as:

$$I_{(A,m)} = P_t|\mathbf{h}_i|^2 \tag{5.6}$$

P_t is the transmission power of the nearby BS that shares the spectrum. The potential impact of all the new user access on a shared spectrum can thus be described as:

$$I' = \sum_{j=1}^{N} I_{(j,m)} \tag{5.7}$$

N is the number of all the secondary users sharing the spectrum through their nearby BSs.

To this end, the procedure of SDN-enabled orchestrated spectrum sharing for user A using 3D interference map could be detailed as follows:

Algorithm 2 Orchestrated spectrum sharing using 3D interference map

User A **moving across potential spectrum area**
User A **sends sensing results and spectrum request to BS**
BS inquires the local SDN controller: spectrum database
if Clear in time domain **then**
    Grant spectrum to user A
  if Out of buffer zone in space domain **then**
    Grant spectrum to user A
    if In buffer zone but $I' \leq threshold$ **then**
      Grant spectrum to user A
    else
      Deny access for $t = t_{thre}$.
    end if
  end if
end if

Note that the execution pre-condition of the orchestrated spectrum sharing algorithm is that a new user sends its sensing results and spectrum request to a nearby BS. Under this scenario, BS inquires local SDN controller for spectrum sharing decision. Local SDN controller then checks the spectrum database: if the spectrum is not used by any incumbents or the new user is out of the protection zone of the incumbents, the controller grants the spectrum for sharing; else, if the new user is within the buffer zone of the incumbents, the 3D interference map is referred, and the spectrum is grant for sharing only when the accumulated interference at the

incumbents I' is within a pre-agreed threshold t_{thre}.

In the proposed orchestrated spectrum sharing scheme using 3D interference map, we utilize the spectrum database that contains the active incumbent location (and geographical area), frequency band and time duration information to facilitate efficient and reliable determination of spectrum availability. Users are free to sense the environment, but their reports only help the local SDN controller in making decisions. To this end, 3D interference map is used to decide the spectrum sharing for a period. If the new access is safe in time and space domain according to the spectrum database, it will be given access directly; else, the cumulated interference at existing users are calculated, and only the secondary access with small potential impacts will be allowed access to protect the incumbents from harmful interference.

5.4 Performance Evaluation

Matlab simulations are conducted to verify the performance of the proposed SDN-enabled orchestrated spectrum sharing compared with traditional methods. We consider a scenario where the users are randomly distributed within a square-shaped area of $1km \times 1km$, and the transmit power of the incumbents is fixed at $40mW$. Assume that there are m existing users working at different frequencies $freq = [f_1, f_2...f_n]$, and n randomly generated new users are requesting access to the available spectrum resources. As PA tier systems get exclusive usage of the spectrum on time/location basis through spectrum bidding, here we only simulate the scenario that how GAA users share the same spectrum with incumbent users and compare the different interference level as well as the ratio of denied access.

Fig. 5.4 shows the simulation setup of incumbents and randomly generated new users who are requesting spectrum resources. Assume that three incumbent users were communicating at frequencies of $900MHz$, $3.5GHz$, and $480MHz$. Here $900MHz$ is licensed spectrum, while $3.5GHz$ and $480MHz$ are incumbent spectrum pools for sharing. $3.5GHz$ is used as one of the spectra for sharing because Federal Communications Commission (FCC) has engaged in creative frequency allocation and recently adopted rules to allow shared commercial use of 150 MHz of spectrum in the 3550-3700 MHz (3.5 GHz) band, called the Citizens Broadband Radio Service [91]. Hence, 3.5 GHz frequency band is selected as an example to show the

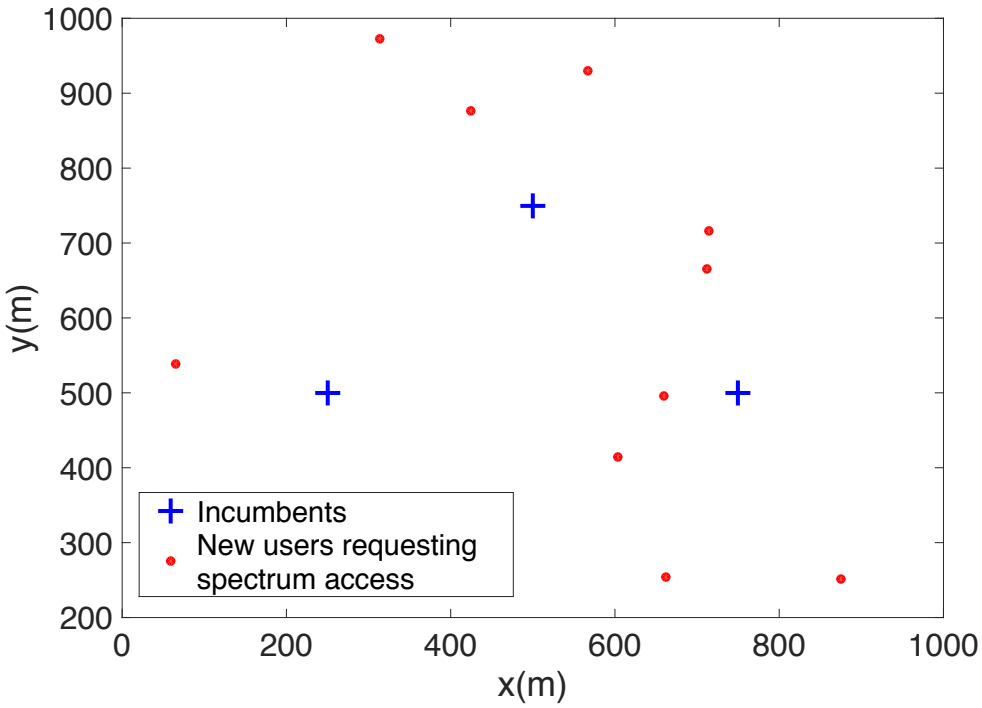

Figure 5.4: Simulation set up of the randomly generated incumbents and new accessing users.

performance, and $480MHz$ is an example for DTV spectrum.

In the simulation of traditional cognitive radio based spectrum sharing method, the new users utilize spectrum sensing based access. The low-complexity energy detection technique is used, and the spectrum is deemed as idle if the received incumbent signal level was less than a predefined threshold. On the other hand, the proposed scheme utilized the spectrum database of the incumbents for estimating the interference imposed on the incumbents by GAA users in different frequency bands. The requested spectrum was allocated to a GAA user if the total sum of the interference at the incumbent imposed by the new transmissions remained below a predefined threshold.

Fig. 5.5 shows the performances of both sensing-based spectrum access and the proposed orchestrated spectrum access schemes regarding the interference imposed on the incumbents. It can be seen that generally, the proposed scheme is lower than the baseline method on the interference that is imposed on incumbents, which means the former scheme has better protection over the existing users. The reason is that in the proposed scheme, BS would first allocate frequency 1, $900MHz$ to the new access, and then inquiry the spectrum database and calcu-

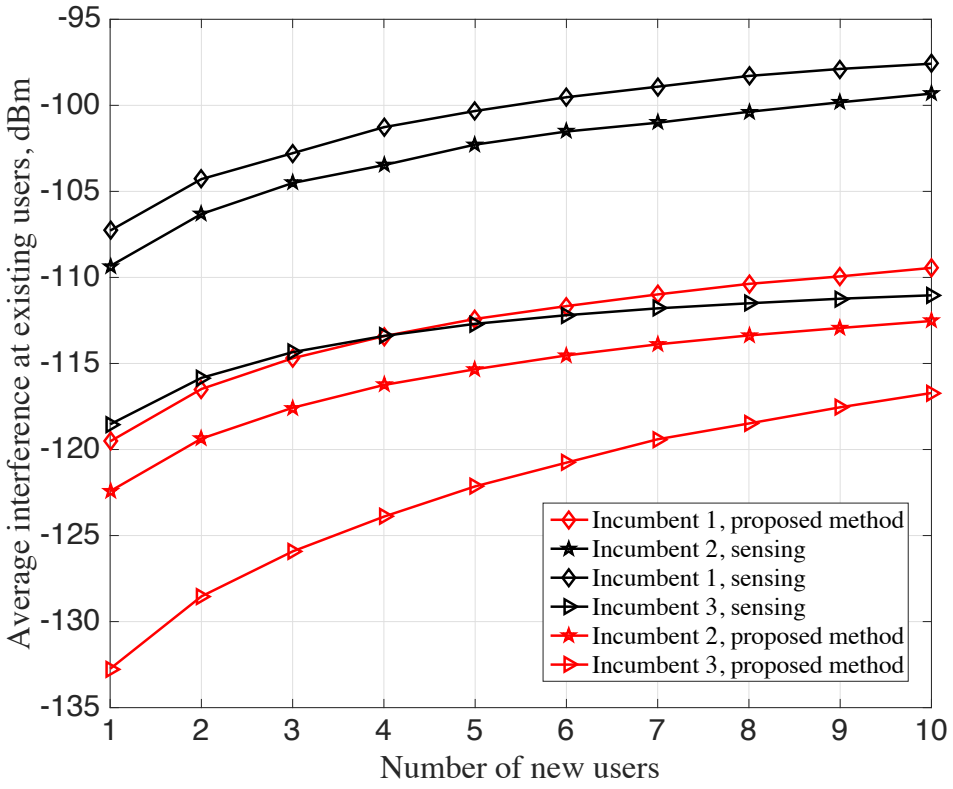

Figure 5.5: The average interference at existing useers in dBm.

late the potential interference map before assigning shared spectrum to GAA users. That's why incumbent frequency 1 has the highest interference. Instead of allowing the devices make decisions themselves, the proposed scheme always try to allocate shared spectrum when the interference at incumbents is tolerable.

Fig. 5.6 compares the performance of the two schemes in terms of the average number of denied access requests (within a single cell of radius $1km$). The average denied access is defined as the ratio of the rejected secondary spectrum access requests to the total number of generated requests. It is clear that the proposed method reduces the proportion of denied spectrum access for GAA users. This is because in the proposed scheme, GAA users are more balanced distributed in the available spectrum, the denied access due to the overload at one frequency thus significantly reduced.

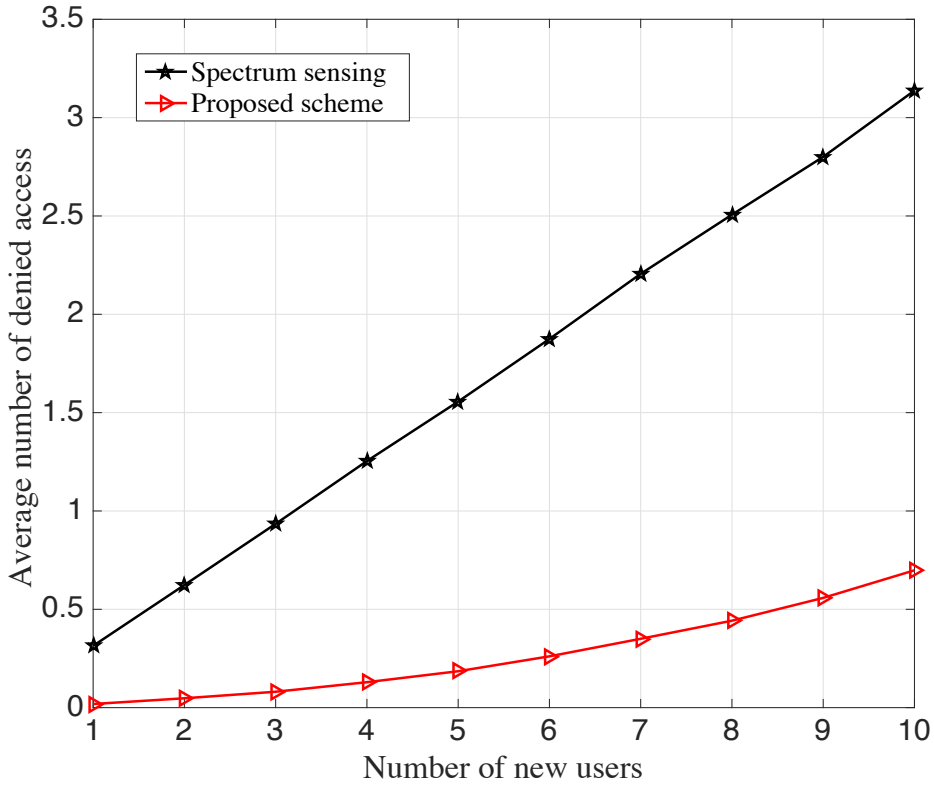

Figure 5.6: The average number of denied access.

5.5 Chapter Summary

Efficient utilization of radio spectrum has been receiving tremendous interest during the past decades. Previous spectrum sharing solutions, especially those based on cognitive spectrum sensing, are prone to inaccurate decisions due to the uncertainties imposed by the wireless medium. Coupled with their heavy reliance on device-level spectrum management, the drawbacks has discouraged network operators from embracing the idea of spectrum sharing. In this article, we proposed an orchestrated spectrum sharing approach that integrates the distributed located UEs/DTV receivers, base stations (BS), DTV stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with real-time information exchange using the return channel of the incumbent system. To adequately protect incumbent DTV users and efficiently share the pooled spectrum resources, real-time 3D interference map is considered to guide the spectrum access based on the SDN global view. Matlab simulation is also conducted and verified the performance improvement.

Chapter 6

SDN-enabled Security Provisioning in 5G HetNets

6.1 Introduction

Over the last decades, anywhere, anytime wireless connectivity has gradually become a reality and resulted in remarkably increased mobile traffic. Mobile data traffic from prevailing smart terminals, multimedia-intensive social applications, and cloud services are predicted to grow at a compound annual rate of 57% before 2019 and is expected to outgrow the capabilities of current fourth generation (4G) and Long Term Evolution (LTE) infrastructure by 2020 [6]. The explosive growth of data traffic is expected to pose a huge burden on the already insufficient spectrum resources in future 5G cellular networks. Therefore, network densification using low-power small cells is widely considered to be a critical solution for 5G driven by the growing influx of mobile data traffic and the in-building wireless coverage requirement [72].

Along with the advantages of this 5G heterogeneous architecture, there arise several significant technical challenges. The massive deployment of small cells poses potential pressure in network management. For example, the reduced cell size induced frequent handover in 5G HetNets could also introduce excessive latency, which could be well beyond the requirements of future 5G applications. Emerging applications supported by 5G, like interactive gaming, real-time and industry-oriented machine communications, require network latency to be even an order of magnitude smaller than 4G [9]. Moreover, the power and resource constraints of

small cell APs require low complexity and high efficient handover authentication procedure, which is hard to realize based on previous cryptographic methods. Therefore, intelligent control over the HetNets for consistent, fast, and reliable handover authentication schemes are still yet to be developed for 5G.

As a mean of non-cryptographic security provisioning method, terminal or user specific physical layer attributes are able to provide a unique fingerprint of the particular device and thus simplify authentication procedure without significant additional hardware and computation cost [79]. Compared with the digital cryptographic authentication method, which brings high risks to the network once security key is spoofed, physical layer attributes are user-inherent and thus hard to be totally compromised. Therefore, user attribute authentication has been studied in many related works. Liu et al. in [92] utilized wireless channel information to generate specific channel impulse response (CIR) for the transmitter-receiver pair. Authors in [93] reviewed physical layer characteristics and analyzed channel state information (CSI) and RSS-based authentication. In [94] and [95], the in-phase/quadrature (I/Q) imbalance and carrier frequency offset (CFO) were used to differentiate the transmitter-receiver pair.

However, completely relying on one physical layer attribute is not deemed as a reliable solution since the selected characteristic may not have enough dynamic range for proper differentiation. For low-security requirement applications like Internet surfing or gaming, quick single attribute verification might be sufficient and reduce complexity. To improve the authentication reliability for high-security requirement applications like banking or online shopping, it is critical to consider more than one physical layer characteristics as secure-context-information (SCI) and verify the claimed device identity in multiple aspects. In [93], multiple physical layer attributes are studied separately for identifying wireless users. This chapter is thereby motivated to provide combined SCI to guarantee the secure level of the authentication as well as reduce authentication latency for 5G HetNet.

In this chapter, we first identify the challenges of security management in 5G HetNets. Based on the observation, we propose a new 5G management structure enabled by Software-Defined Networking (SDN), to bring intelligence and programmability into 5G HetNets for efficient security management. With SDN, the control logic is removed from the underlying infrastructures to a controller in control layer [21]. The software can then be implemented on

the central SDN controller to provide consistent and efficient management over the whole 5G HetNets. With this paradigm, we propose an SDN-enabled fast authentication scheme using SCI transfer to achieve seamless authentication during frequent handovers, while at the same time meet the latency requirements.

The remainder of this chapter is organized as follows: Section 6.2, 6.3 and 6.4 presents the SDN-enabled fast authentication algorithm, including SDN system model and weighted SCI design. SDN-enabled privacy protection scheme is given in Section 6.5. Further, performance evaluation is presented in Section 6.6. Finally, the conclusion is drawn in Section 6.7 .

6.2 System Model

6.2.1 SDN-enabled HetNets Model

SDN is introduced in order to enable the coordination between HetNets, as shown in Fig. 6.1. Authentication handover module (AHM) is implemented into the SDN controller to monitor and predict the location of users, and then prepare the relevant cells before the user arrives to guarantee seamless handover authentication. Using traffic flow template (TFT) filter [58] (source/destination IP addresses and port numbers) and related quality of service (QoS) description, SCI is collected by AHM to share along the user projected moving path, i.e., from cell *A* to cell *B*, *C* and *D* in Fig. 6.1. The relevant cell APs thus prepare resource in advance and ensure seamless user experience during mobility.

More precisely, the way that SDN controller shares the user's SCI to next cell APs along the predicted path is just like a trustworthy introduction from previous AP before the handover. The future cell APs thus finish authentication with the user quickly and begin to monitor the user to prepare service according to the SCI. As the trace of the user is controlled, the risk of impersonation is significant, if not entirely, reduced. More importantly, under the condition of service disruption when the connection between APs and the authentication server is broken, the proposed mechanism will not lose global network connectivity because a new AP is monitoring the user, which can help the controller to retrieve the necessary information according to the pre-shared SCI. Thus, the SDN-enabled security handover possesses high levels of

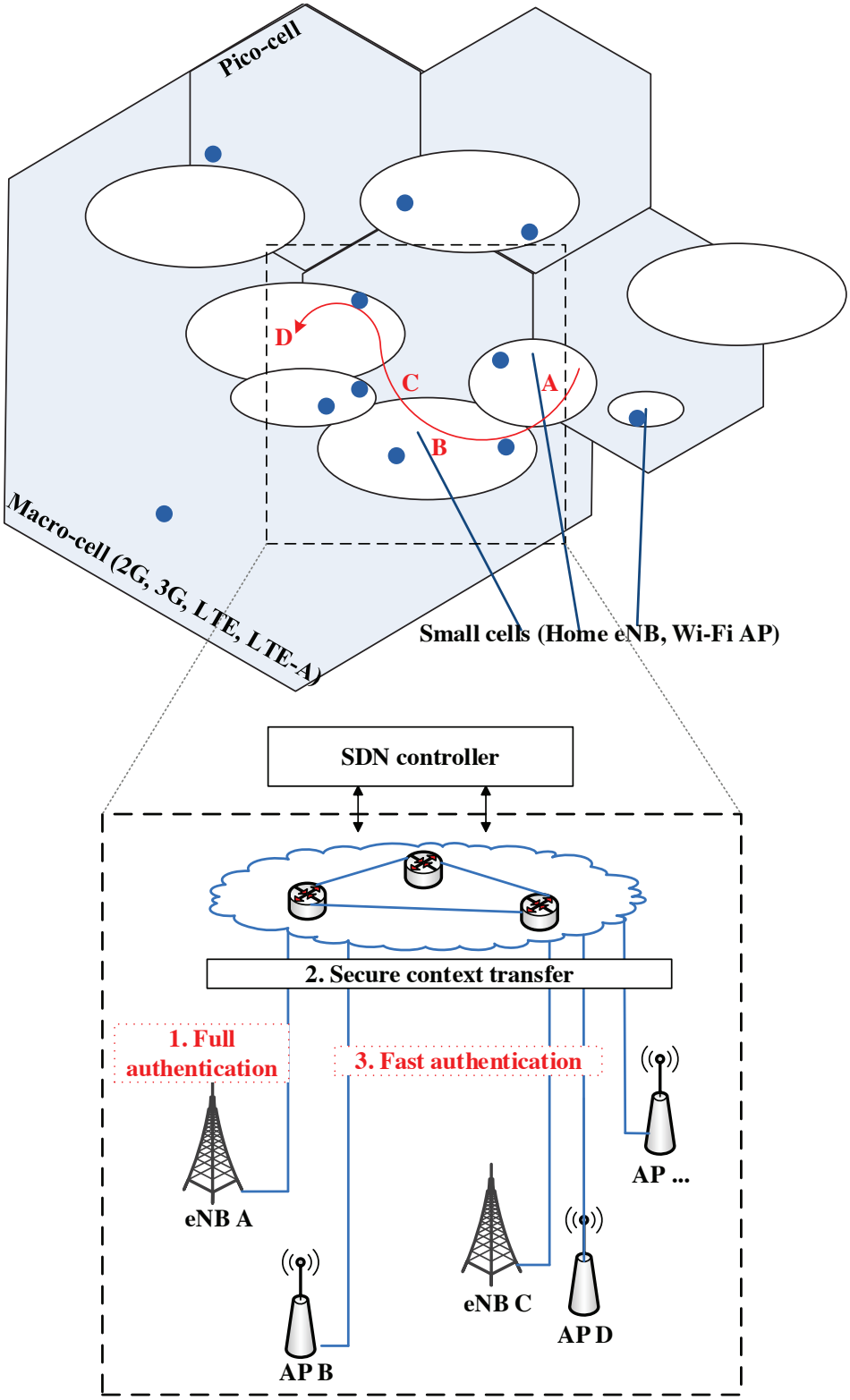

Figure 6.1: SDN-enabled secure context information transfer between 5G UE, APs and AHM in SDN controller.

tolerance to network failures.

The system model for SCI-based authentication using unique user physical layer attributes associated with each transmitter-and-receiver pair is illustrated in Fig. 6.2. The commonly used terminology of Alice, Bob and Eve is applied here to represent legitimate transmitter, receiver, and eavesdropper. Alice sends a message to Bob, while Eve tries to mimic Alice during some communication time slots. Therefore, it is clear that Bob needs to be able to discriminate Alice from Eve using pre-shared or pre-defined "keys", which is the combination of physical layer attributes obtained from the received signals in our proposed non-cryptographic authentication method.

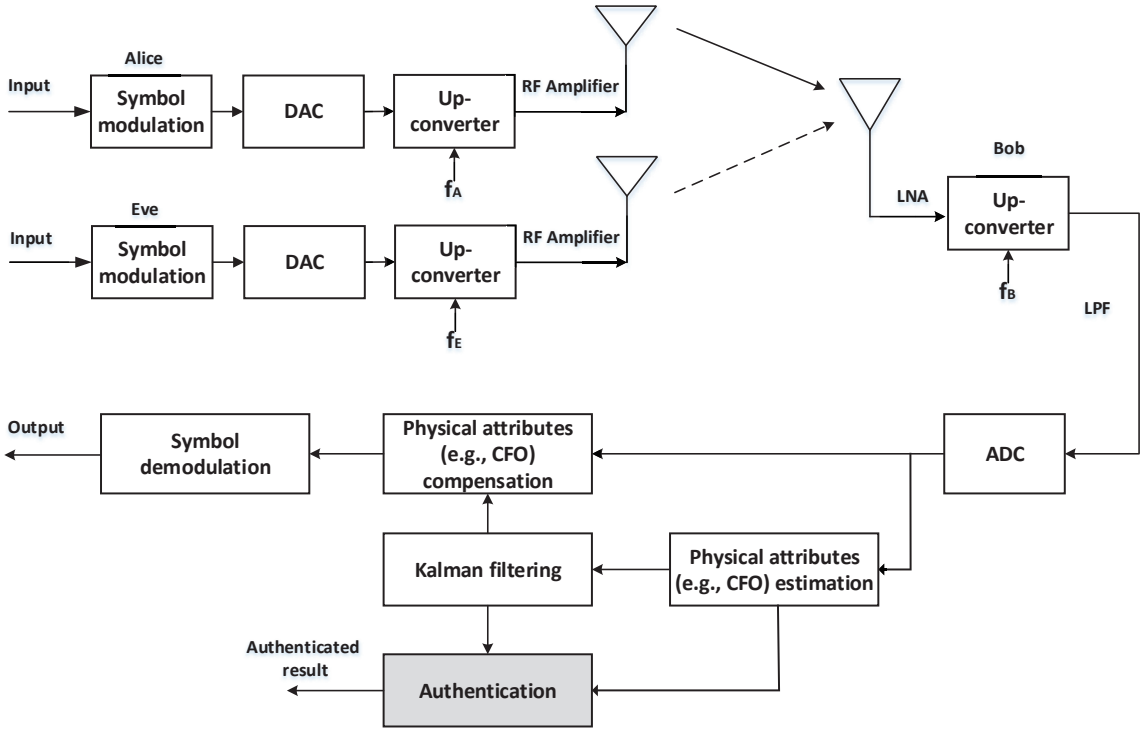

Figure 6.2: SCI based authentication using unique user physical layer attributes associated with each transmitter-and-receiver pair.

As shown in Fig. 6.2, Alice generates baseband signals, which is then modulated to a carrier frequency f_A after digital-to-analog conversion (DAC). The signal is then transmitted and arrived at Bob. The radiated RF signal would then be converted to the baseband. As there are different oscillator characteristics and Doppler shift due to mobility, some physical layer attributes would be unique between Alice and Bob. For example, the carrier frequency

of Alice and Bob would not be the same, resulting in a carrier frequency offset (CFO), i.e., $\delta f_A = f_A - f_B$, while the CFO between Eve and Bob would be $\delta f_E = f_E - f_B$. Due to the manufacture independence of devices, the received signal that is sent from different transmitters would experience distinctive CFOs at the various time instant. Therefore, the unique physical layer attributes such as CFO that are associated with each pair of transmitter and receiver can be used as the identification if combined properly.

6.2.2 SDN-enabled Fast Authentication Model

6.2.2.1 Assumptions and design goals

We assume that SDN controller is a program running in mobile operator's data center with an authentication handover module (AHM) for user authorization. The AHM is in charge of both authentication and handover, which maintains user information specifying what the user can access. The AHM also possesses a master public-private key pair (K, K^{-1}), with a public key K that is known to users and APs. Both APs and UEs need to be verified before getting access to network services to reduce security risks.

Our design goal of the authentication handover mechanism is to accelerate authentication procedure in 5G HetNet by enabling SCI transfer using SDN. In further reducing the overall authentication delay, AHM in the controller could periodically authenticate the APs in off-peak times using its master key to avoid leakage of privacy caused by compromised APs. If certified, a key pair (K_N,K_N^{-1}) with a signature$[K_N, T]_{K^{-1}}$ is distributed to the AP, where T is the timeout of the signature; if the AP is detected as compromised, it will be blacked out from further operation. This way, part of the authentication procedures are moved to off-peak times and relieves SDN controller burden.

6.2.2.2 Fast authentication mechanism design

With the assumptions and design goals described above, we can develop the SDN-enabled fast authentication mechanism. User specific SCI, such as ID, physical layer attributes, location, speed, and direction, can be collected and shared easily with SDN flow based forwarding [21]. According to the UE location information from SCI, SDN controller uses ascending index to

indicate the sequential order of next cells in the moving direction. Once authenticated by one cell AP, an appropriate combination of user attributes is then shared as SCI by SDN controller along this user's future path for a valid time duration. This way, the UE can enjoy seamless service without complex operation during that periods, thus reducing latency for data communications.

For example, we assume that a user U is in cell A is heading to cell B and C, as shown in Fig. 6.1. The authentication procedure between user U and cell A follows the commonly used authentication protocol [48], and the proposed SDN-enabled fast authentication procedure is described as follows:

Algorithm 3 User SCI based fast authentication procedure

0: **State(A, U):** Authenticated.
State(B, U): Not Authenticated.
State(C, U): Not Authenticated.
AHM → B: $(index = 1, ID, SCI)$
0: **AHM → C:** $(index = 2, ID, SCI)$
Ascending index number shows the direction of user movement. ID is the identity of U and SCI is the secure context information of U.
0: **B → A:** Handoff REQ(ID, SCI).
When B discovers U in its coverage, B sends handoff request to A until receives reply from A.
0: **A → B:** Handoff ACK$(ID, SCI^{'})$.
A replies with handoff acknowledgement. $SCI^{'}$ is the secure context information which is more recent than previous shared SCI.
0: **B → U:** Update REQ().
After matching $SCI^{'}$ from A with U, B authenticates U and starts to associate with U.
0: **U → B:** Update ACK$(SCI^{''})$.
Here U is connected with B. $SCI^{''}$ is the latest secure context information.
State(B,U): Authenticated.
0: **B → AHM:** Update$(SCI^{''})$.
B updates the UE secure context information to AHM. AHM then shares secure information to next cell APs according to the location and direction information in new $SCI^{''}$.
0: **C → B:** C keeps on monitoring U and follows similar procedure.

6.2.2.3 Physical layer attributes

The SCI attributes in the proposed SDN-enabled authentication handover could include identity, physical layer attributes, location, moving speed and direction. Take one of the well-known physical layer characteristics, carrier frequency offset (CFO) as an example, CFO can be utilized as a radiometric signature for wireless device authentication because that radio frequency

oscillators in each transmitter-and-receiver pair always present device-dependent biases to the nominal oscillating frequency in realistic scenarios [95]. Therefore, the combination of these bias and mobility-induced Doppler shift constitute the variable CFO values at different communication time. CFO can be estimated first and compared with the current value to determine whether the signal has followed a consistent pattern and thus can be seen as legal.

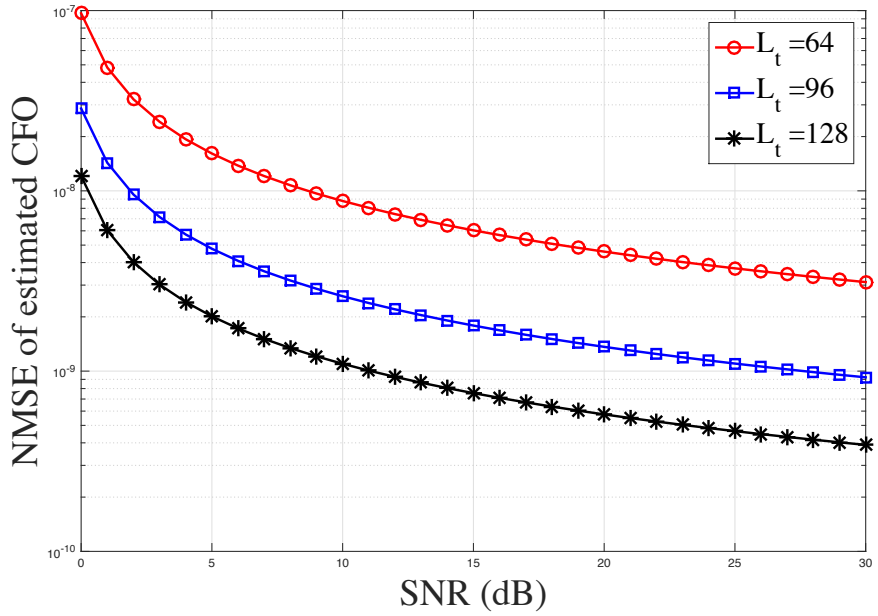

Figure 6.3: NMSE of CFO estimates vs. SNR with different training lengths.

Fig. 6.3 shows the normalized mean square error (NMSE) of the estimated CFO versus SNR. It can be seen that the estimation error of CFO decreases with the increase of SNR. This is reasonable because estimation would be more accurate with better channel quality. It can also be observed that the estimation error reduced when the length of the training segment increases.

The CFO based authentication scheme works as follows: firstly, the current CFO estimate $\hat{\epsilon}[m]$ and the corresponding mean sqare error (MSE) matrix M_0^0; secondly, the predicted CFO vector and prediction MSE are obtained [95]; thirdly, the decision threshold $T[m]$ is derived using channel condition parameters, using:

$$T[m] = \sqrt{\frac{\eta}{\gamma[m]} + M_{M=1}^{M}(2,2)}Q^{-1}(\frac{p_f}{2}) \tag{6.1}$$

where η is a system variance, i.e., $\eta = 1/(4\pi^2 L_t^3)$, L_t is the length of the training sequence.

$\gamma[m] = \eta/\sigma^2[m]$, where $\sigma^2[m]$ is the estimation noise variance of the mth data frame.

Finally, the absolute difference between CFO estimate and the predicted CFO are compared with the decision threshold $T[m]$ to see whether the hypothesis is valid. If the difference is smaller than the threshold $T[m]$, the CFO state is updated and goes to next frame; else, the current frame is considered from an illegitimate transmitter, i.e., Eve, and the transmission requests will be rejected.

Fig. 6.4 illustrates how the threshold T changes with the increase of SNR and false alarm rates P_f. It is observed that the decision threshold becomes smaller when the SNR increases, which is due to the fact that a better estimation performance is achieved with good channel quality, and thus the difference between the estimation and received value would be small. For a fixed SNR, the threshold value increases as the corresponding false alarm rate decreases, which is also reasonable as lower false alarm rate require a larger confidence interval to accommodate the estimation error.

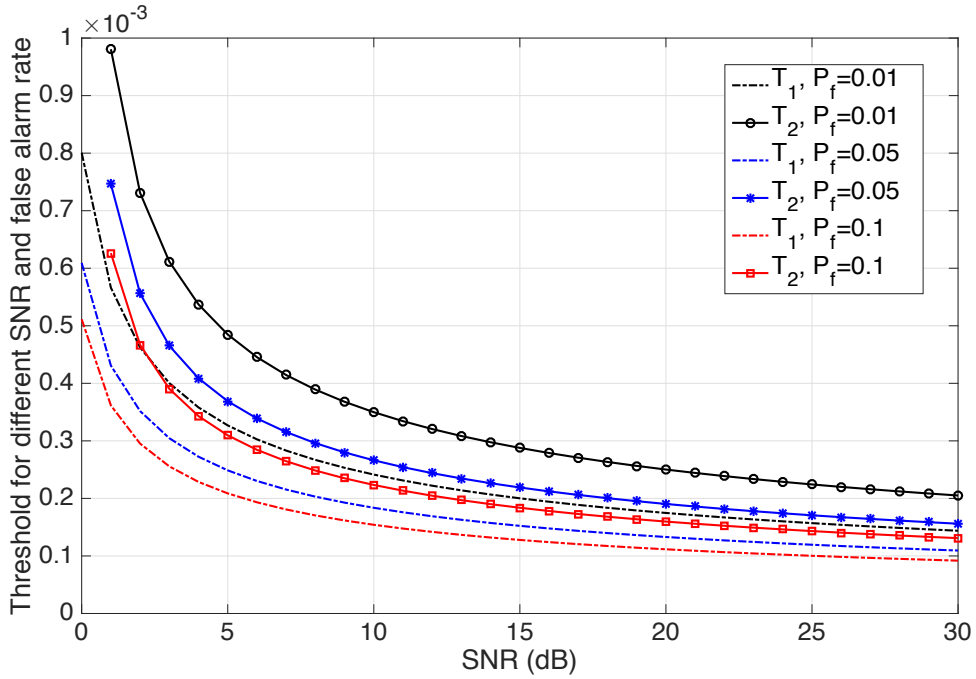

Figure 6.4: Decision threshold for CFO versus SNR for different false alarm rates P_f

There are different kinds of physical layer attributes. Next, location and received signal strength information (RSSI) will be discussed as another example. However, the results are expandable.

Mathematically, the RSSI at receiver from one wireless transmitter is defined as:

$$P(d) = P_{(t)} - 10\alpha log_{10}(d/t) + W \tag{6.2}$$

where d is the distance between the wireless node and receiver, $p_{(t)}$ represents the transmission power. α is the path loss exponent, and N means zero-mean Gaussian noise.

Regarding location, a simple reference node (Here noted as APs) based localization method is adopted to accommodate with the simplified APs in 5G small cells. It is assumed that there is a localization list in SDN controller, which maintains the following fields for UE location estimation: last beacon time (LBT), UE speed in the direction of the x-axis (S_x) and y-axis (S_y) [96]. When the UE get access to a new AP, the AP receives a new beacon and the current time is stored in LBT together with the location of the UE. Afterward, SDN searches its localization list for previous beacon information from the same UE. If previous beacon information from the same UE is found in the list, current UE speed, S, is calculated accordingly.

Specifically, the previous location and beacon time of a UE stored in the list is denoted by (x_1, y_1, PBT) and the same information found in the last beacon packet for the same UE is denoted by (x_2, y_2, LBT). The current UE speed S can be derived according to S_x and S_y [96]:

$$S_x = (x_2 - x_1)/(LBT - PBT) \tag{6.3}$$

$$S_y = (y_2 - y_1)/(LBT - PBT) \tag{6.4}$$

$$S = \sqrt{S_x^2 + S_y^2}. \tag{6.5}$$

The current location region of a given UE (X_{est}, Y_{est} is thus estimated based on the calculated speed and the amount of time passed since LBT whenever a UE is trying to acquire service from the APs. This region is further defined by a circle which centers at (x_2, y_2) with radius S. In the realistic scenario, SCI files should consist of UE ID, location, velocity, the day of the week, MAC address, RSSI, clock skew, I/Q imbalance.... Although usually users tend to keep the similar speed and daily routine, due to the possibility of transportation tools changing, there could be a lowest and a highest threshold in determining the radius in practical and

realistic scenarios. Our linear location prediction scheme is simple, but very reasonable when the beacon interval and the time since LBT are reasonably small, which is just the case in 5G small cells.

Note that in the proposed SCI-based authentication handover scheme which uses a combination of multiple physical layer attributes to verify a transmitter, the number of attributes to be used relies on the security level of the information requested. For example, if the user is seeking for banking or email services, the higher security level can be achieved by transferring more SCI attributes; else if it is just Internet browsing or video, gaming, the security level can be lower, and few SCI attributes are needed.

The aforementioned fast authentication method requires no changes to the existing UE and AP hardware, while significantly simplifies the authentication procedure and reduces handover latency through non-cryptographic technique. By predicting user moving path and shifting the authentication of APs to off-peak times, the SDN-enabled 5G HetNets can always be well prepared for other service requests. Moreover, operators can choose to switch off/on low loaded cells if the users approaching these cells are not going to exceed a certain threshold according to the SCI information to save more energy.

6.3 Weighted SCI Design and Decision Rules

There are three kinds of fingerprint for mobile UEs, i.e., software-based, hardware-based and channel/location-based characteristics [93]. Software-based fingerprinting includes MAC layer behaviors, such as probe request, frame sequence number. Hardware-based character means radiometric fingerprinting, like clock skew and physically unclonable functions (frequency error, I/Q offset, magnitude error and phase error) [97], while channel/location-based fingerprinting means channel state information (CSI) and received signal strength (RSS). CSI and RSS are location-specific due to path loss and channel fading.

User-specific attributes that we consider in this article are a proper combination of the aforementioned physical layer characteristics. Such attributes have been taken as reliable SCI to assist fast and secure handover in 5G networks, instead of using complex and forgeable cryptographic exchange mechanism [93]. Assume that SDN controller obtains the original

file of the user-specific attributes after a full authentication, which shows $X = [X_l, X_q ... X_c]^T$, $\sigma^2 = [\sigma_l^2, \sigma_q^2 ... \sigma_c^2]^T$. Here $l, q, ...c$ corresponds to user attributes like location, I/Q imbalance,... channel impulse response (CIR), and X_l, X_q, X_c are the mean value of the attributes, respectively. Similarly, σ_l^2, σ_q^2 and σ_c^2 represents the variance of the chosen attributes.

In order to show the performance of the multiple SCI combination, SDN controller constantly samples multiple physical layer attributes from the received packets and each time sample a certain number of N packets. The attribute mean and variance are then calculated accordingly for verification. To be specific, the matrix below shows the received file at the receiver, which is a $N * M$ matrix. Here N is the number of attributes, and M means each attribute has M observations at the receiver:

$$\begin{bmatrix} x_{l1} & x_{l2} & \dots & x_{lM} \\ x_{q1} & x_{q2} & \dots & x_{qM} \\ x_{c1} & x_{c2} & \dots & x_{cM} \\ \dots & \dots & & \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}$$

After receiving the above observations, we further define $X' = [X'_l, X'_q ... X'_N]^T$ as the mean of the current sampled N packets, and $\sigma^{2'} = [\sigma_l^{2'}, \sigma_q^{2'} ... \sigma_N^{2'}]^T$ as the variance of the current sampled N packets.

Firstly, we compare the observed attributes matrix through zero-mean white Gaussian noises (AWGN) with the validated original attributes matrix X, and obtain the attributes offset as $\triangle X = X - X'$, expressed as: $\triangle X = [\triangle X_l, \triangle X_q ... \triangle X_N]^T$.

Secondly, a decision rule is needed to successfully determine the eligibility of the UE, namely, authenticate the user. To add flexibility into the proposed algorithm and be able to give higher weight to the attributes that are hard to be spoofed, each attribute will be compared with the threshold, and the final decision is made according to a weighting of the results[94].

In the weighted SCI transfer based fast authentication method, the two-variable mean-shifted hypothesis testing model can be defined as follows:

$$H_0 : A = W \tag{6.6}$$
$$H_1 : A = \triangle X + W$$

Precisely, if the difference between the original and received matrix, $\triangle X$, is less than a predetermined threshold, H_0 is accepted, which means that the UE is legitimate; otherwise, H_1 is decided. To be specific, the joint PDF of the mth attributes under H_0 can be expressed as in [98]

$$P_n(A; H_0) = \frac{exp[-\frac{1}{2\sigma_n^2}\sum_{i=0}^{M-1} x_{[n][i]}^2]}{(2\pi\sigma_n^2)^{M/2}} \tag{6.7}$$

where $n = 0, 1, 2...$ refers to the chosen attributes. Similarly, the joint PDF under H_1 is:

$$P_n(A; \Theta_1, H_1) = \frac{exp[-\frac{1}{2\sigma_n^{2'}}\sum_{i=0}^{M-1}(x_{[n][i]} - \triangle X_n)^2]}{(2\pi\sigma_n^{2'})^{M/2}} \tag{6.8}$$

here $\Theta_1 = [\triangle X_n, \sigma_n^{2'}]$ under H_1. In practice, $\sigma_n^{2'}$ is calculated according to the sampling results, while $\triangle X_n$ is obtained using X minus the sampling mean.

Next, the decision rule of the hypothesis testing procedure is presented to realize the authentication. The decision is based on if

$$L_{n(x)} = \frac{P_n(A; \Theta_1, H_1)}{P_n(A; H_0)} > \gamma \tag{6.9}$$

where γ is a threshold, which decides whether H_0 or H_1 is accepted.

Bringing (6.7) (6.8) into (6.9), one can obtain the threshold for the attributes as below:

$$T[n] = \sigma_n^2 \sqrt{|\gamma^{2/M} - 1|} \tag{6.10}$$

where $n = 0, 1, ...N$. $T[n]$ is the threshold for attributes, e.g., CIR or CSI. To be specific, the authentication of each variable is based on the comparison between $\triangle X$ and $T[n]$.

The false alarm rate of each attribute can thus be derived as:

$$\begin{aligned} P_{F[n]} &= P\{\triangle X[n] > T[n]|H_0\} \\ &= \int_{T[n]}^{\infty} P_n(A; H_0)dS \\ &= 2Q(\frac{T[n]\sqrt{M}}{\sigma_n}) \end{aligned} \tag{6.11}$$

Accordingly, we get the detection probability.

$$\begin{aligned} P_{D[n]} &= P\{\triangle X[n] > T[n]|H_1\} \\ &= Q(\frac{T[n] - \triangle X_n}{\sqrt{\sigma_n^2/M}}) + Q(\frac{T[n] + \triangle X_n}{\sqrt{\sigma_n^2/M}}) \end{aligned} \tag{6.12}$$

In the decision rule, the final decision is based on the majority variables' decision. Given that the multiple variables may claim different results, the final decision rule should take this case into consideration. Precisely, if one variable claims Alice while the others claim Eve, we consider that the received packets are transmitted by Eve. In this way, the performance in detecting spoofing attacks will be enhanced since it is nearly impossible to spoof Bob in multiple aspects, like CIR, packet error rate, and even location at the same time.

6.4 SDN-enabled Fast Authentication Algorithm using Weighted SCI Transfer

In the following, a description of the detailed SDN-enabled fast authentication mechanism is presented. To adapt to the simplified 5G small cell APs, the fast authentication scheme should be defined as simple as possible. User-specific SCI is, therefore, pre-transferred to future APs by SDN controller for prepared and fast authentication. In doing so, a neighbor graph of HetNet is formed by SDN to predict the projected cells along the moving direction according to the collected SCI, which means cell B, C and D in Fig. 6.1. The proposed fast authentication protocol includes two parts: weighted SCI transfer based Fast Authentication and Full Authentication (as in 802.11), as shown in Algorithm 4 below.

Algorithm 4 SDN-enabled fast authentication using weighted SCI transfer

First time arrived:
Full authentication; SCI sent to AHM and shared along the moving path with a valid duration t_v
if $t \leq t_v$ **then**
  Execute Weighted SCI transfer based Fast Authentication
else if t_v time out **then**
  go back to second step: Full authentication; SCI sent to AHM and shared with another valid duration t_v
end if

The fast authentication procedure mentioned in the Algorithm works as follows: after the first Full Authentication in cell A (in Fig. 6.1), once the UE is arriving at a new cell B, C or D, SDN controller collects the secure level according to its traffic descriptions. If its secure level is low, only single attribute (e.g., MAC address) will be verified; otherwise if there are high secure level operations, combined weighted SCI verification (introduced in the last section) is executed to guarantee security. MAC address verification only needs local processing, so the user with normal operations can handover seamlessly without too much authentication delay. Moreover, fast authentication can even be done in the background while the user continues to communicate. In this way, the processing time is shortened considerably because weighted SCI based fast authentication occurs locally and simply.

Furthermore, if a carrier regards SCI-only user authentication is not secure enough, the valid duration t_v is free to be set smaller. The proposed scheme not only reduces authentication delay but also provides considerable flexibility, and thus supports seamless inter-AP mobility without a significant sacrifice of secrecy in practical and realistic scenarios.

6.5 SDN-enabled 5G Privacy Protection

Data privacy means the right for network users to seclude themselves from prying and eavesdropping. Due to the reduced cell size in 5G HetNets, users might move through multiple small cells before completing one communication session. The privacy protection thus is more challenging in 5G due to the possible involvement of untrusted or compromised APs during handover. Existing privacy protection schemes use complex key agreements and interactions

or new watermark to protect data privacy. Such cryptographic methods bring computation burden and complexity to both AP and client side [2], which is undesirable for 5G low power small cell infrastructures. On the other hand, privacy protection requires that no link could be established between information and the owner, while authentication requires an identity provided for the purpose of authenticating. Previously these contradictory requirements are met through a trusted third party. However, multiple inquiries to the remote third party bring network bottleneck, which is not suitable for 5G low latency communications.

We introduce SDN-enabled privacy protection scheme, which employs partial transmission over different SDN controlled network paths to guarantee privacy and offload traffic from cellular networks at the same time. With the proposed privacy protection scheme, SDN controller is able to choose multiple network paths to transmit different part of the data stream, i.e., partial transmission, according to the heterogeneous network coverage. The number of network paths is decided by the sensitive level of the data stream. As long as the UE has been authenticated and is covered by the heterogeneous networks, e.g., Wi-Fi, Femtocell or cellular, the induced data stream can be routed through these network backhauls under the control of SDN controller. Only the receiver can decrypt the data using its private key and then re-organize the data stream coming from multiple network paths, which avoids privacy leakage by compromised APs. Moreover, the proposed scheme can realize traffic offloading through the other network paths, which is desirable given the fact that 5G cellular network would be flooded by a sheer volume of mobile traffic [6]. Simply by choosing nearby Wi-Fi or Femtocell as the different paths for data offloading, traffic load of the 5G cellular network is relieved through either unlicensed band of Wi-Fi or reused band of Femtocell. The proposed SDN-enabled privacy protection mechanism is described below:

Algorithm 5 SDN-enabled privacy protection algorithm using multi-path transmission

1: **Delay threshold**:T_s
2: **Number of available networks**:n
3: **Size in bytes to be transferred in nearby Wi-Fi, Femtocell or cellular within** T_s:$V_{sn} = b_n\ min(t_r, T_s)$
4: **for** $d_1 < V_{s1}$, $d_2 < V_{s2}$, $...d_n < V_{sn}$ and $d = d_1 + d_2 + ... + d_n$ **do**
5: *Encrypt $d_1, d_2, ..., d_n$ separately, send them on n networks concurrently and update d*
6: **end for**
7: *Receiver decrypt $d_1 \sim d_n$ using private key and re-organize data*

Here n is the number of network paths that SDN controller chooses for data transmission, d_n is the different part of data that will be transmitted in nth network concurrently. t_r is the data transfer time within the involved networks. T_s is the delay threshold of 5G applications, which means that this kind of service needs to be finished before T_s to guarantee user experience. For example, email transfer can tolerate long latency, while real-time video, two-way gaming has a very low delay threshold. b_n is the bandwidth allocated by SDN controller according to the traffic situation of different networks and V_{sn} is the volume of data that can be transferred in the multiple paths, i.e., offloading networks, within the application delay threshold.

More importantly, the number of paths n here is decided by a trade-off between privacy level, offloading revenue, and system complexity, which is reconfigurable and can be easily set up through SDN controller application by 5G operators. User privacy protection thus becomes programmable and under control of SDN, which is especially desirable for future high diverse communication requirements and application QoS needs.

6.6 Performance Evaluation

6.6.1 SDN Modeling using Priority Queuing

In order to measure the performance of SDN-enabled fast authentication, we need to model SDN properly. Related works in [61] et al. provided SDN modeling, however, their calculation only based on Poisson arrivals. Moreover, they failed to take the higher priority of Controller processing over switches into consideration. Therefore in this work, we propose a new SDN network model using priority queuing, and model the arriving Internet traffic as Pareto distribution [62]. Figure 6.5 shows our queuing model for the SDN controller and switch.

In Figure 6.5, λ is the data arrival rate while μ_s and μ_c represent the processing rates at the switch and the controller, respectively. The processing time of SDN-enabled network depends on the flow table within the switch, i.e., whether or not the switch's flow table contains a rule for the incoming traffic flow. As shown in the figure, if the packet arriving at the switch is the first packet of a new data flow (new source-destination pair), the switch forwards this packet to the controller. The controller decides the optimal forwarding rule for this packet and returns it

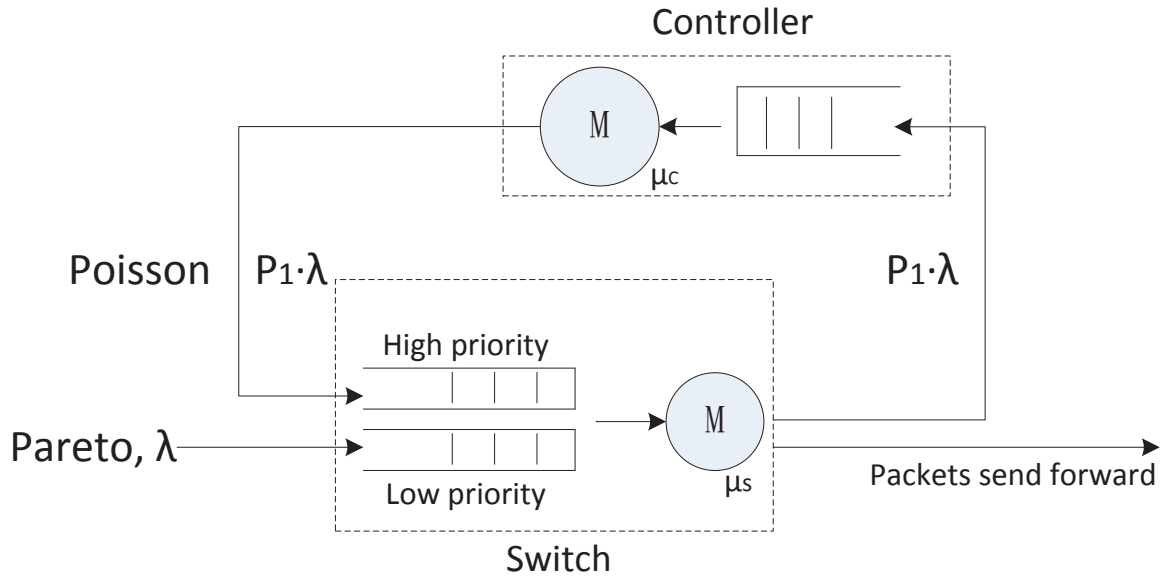

Figure 6.5: A model for SDN switch and controller.

to the switch [79].

Controller processing has higher priority over switch forwarding because the simplified OpenFlow switches always follow the controller instructions. Therefore, the High priority queue is always served first (See Figure 6.5); then, if the High priority queue is empty, the other queue is processed. Anytime a packet is received in the High priority queue, that queue is emptied before the other queue is processed. In the meanwhile, the controller processing also includes pushing the corresponding forwarding rule into the flow table of the related switches. Subsequent packets from the same flow are forwarded based on this newly installed forwarding rule. It is worth mentioning here that the following analysis relies on the assumption that the incoming traffic is TCP, i.e., a given source node initially transmits a single packet to initiate the TCP handshaking procedure, and the actual data shall be forwarded once the TCP session is established.

Assuming P_1 represents the probability that there is no flow entry in the Openflow switch for the incoming packet, the average packet delay, T_D, can be written as in [79]

$$T_D = T_{D1} \times P_1 + T_{D2} \times (1 - P_1), \tag{6.13}$$

where T_{D1} is the delay incurred if the switch has to forward the packet to the controller

while T_{D2} represents the delay which occurs when the switch forwards the data directly, i.e., a forwarding rule for the packet already exists. It has been shown in [61] that in a normal productive network carrying end-user traffic, a switch observes new flows with a probability of 0.04, hence, we use this value for the probability P_1.

The delay T_{D1} can be further written as

$$T_{D1} = T_C + 2T_{PROP} + 2T_{sw}, \tag{6.14}$$

where T_{sw} and T_C, respectively, represent the delays at the switch and the controller, including the queuing and the processing delays. Moreover, T_{PROP} denotes the propagation delay between the controller and the switch. Note that T_{sw} in (6.14) is multiplied by two since the packet has to make two passes through the switch. In the first pass, it is forwarded to the controller for further processing, and in the second pass, it is forwarded along the data path once a forwarding rule has been established. On the other hand, $2T_{PROP}$ accounts for the propagation delays when the packet is sent to and from the controller. The delay T_{D2} is equal to switch processing time only, i.e.,

$$T_{D2} = T_{sw}. \tag{6.15}$$

Next we derive the delays T_{sw} and T_C. Here, it is worth to mention that the packets returning from the controller are Poisson distributed, while the arriving Internet traffic is Pareto-distributed. These two flows of traffic arriving at the switch follow the priority queuing system.

T_C: The arrival rate at the controller is Poisson distributed since the departure rate at the switch is Poisson [66]. Therefore, T_C can be calculated using the waiting time equation for $M/M/1$ queuing as

$$T_C = \frac{1}{\mu_c(1-\rho_c)}. \tag{6.16}$$

In the above equation, ρ_c represents the controller utilization and it is equal to $P_1 \cdot \lambda/\mu_c$ (See Figure 6.5).

T_{sw}: In order to derive T_{sw}, the Internet arrival data at the switch is modeled as Pareto distribution, with shape parameter α and scale parameter k [62]. Furthermore, the switch processing

times are modeled as exponentially distributed with rate parameter μ_s. With the aforementioned assumption and the priority queuing model, the average waiting time E_{W_p} for a new packet is given as [99]

$$\begin{aligned} E_{W_p} &= \frac{E_V}{(1-\rho_c)(1-\rho_c-\rho_s)} \\ &= \frac{E_V}{(1-P_1\lambda/\mu_c)(1-P_1\lambda/\mu_c-\rho_s)} \end{aligned} \tag{6.17}$$

Here ρ_c is the utilization rate of controller processing queue, i.e., High priority queue, and ρ_s relates to switch processing queue. E_V is the residual time of the jobs in service.

Recall that packets from the Internet are Pareto distributed with arrival rate λ, and packets after controller processing are Poisson distributed with arrival rate $P_1\lambda$. Consider the data file d to be transmitted as Pareto distribution and according to the CDF of Pareto distribution $F_d(X)$, we further get

$$F_d(D) = 1 - (\frac{k}{D})^{\alpha} \; for \; D \geq k \tag{6.18}$$

$$\rho_s = \frac{\lambda}{\mu_s} = \frac{1}{E(d)\mu_s} = \frac{\alpha-1}{\alpha K \mu_s}. \tag{6.19}$$

Next, we try to derive the residual time of the job in service E_V. According to priority queuing theory [99], one get

$$E_V = \sum_{k=1}^{2} \frac{\lambda_k(E[X_k^2])}{2} = \frac{2}{\mu_s^2}(\frac{P_1\lambda}{2} + \frac{\lambda}{2}) \tag{6.20}$$

where X is the service time. Apply (6.20) into the packet average waiting time E_{W_p} of the priority queue at the switch (6.17), we further get

$$E_{W_p} = \frac{P_1\lambda + \lambda}{{\mu_s}^2(1-P_1\lambda/\mu_c)(1-P_1\lambda/\mu_c-\rho_s)} \tag{6.21}$$

Finally, we achieve

$$T_{sw} = E_{W_p} + \frac{1}{\mu_s}. \tag{6.22}$$

Using (6.14)-(6.22) in (6.13), one can find the average packet delay with SDN based data forwarding.

6.6.2 Performance Evaluation of Proposed Fast Authentication Algorithm

6.6.2.1 SDN network latency

MATLAB simulations of the 5G network with commonly used hexagonal cells are adopted to evaluate the performance of the aforementioned mechanisms regarding the secure level and latency. A total of 19 small cells in Fig. 6.6 with an inter-site distance (i.e., the distance between two AP) of 300m is considered in the simulation. Users are randomly distributed around APs, while each UE engages a random walk and changes direction every 5 seconds. Wrap-around technique (i.e., users move out of the pre-defined service area are assumed to enter the area from the other side of the network) is used to avoid boundary effects. The specific simulation parameters are listed in Table 1.

Table 1: Simulation parameters of 5G networks.

Cell layout	Hexagonal grid, 19 cell sites, with wrap-around technique
Cell radius	150m
User mobility speed	3km/h
User mobility direction	random
Total number of users	570

In order to evaluate the latency reduction performance of the proposed non-cryptographic algorithm compared with traditional cryptographic methods, SDN controller processing ability is simulated in Matlab. Without loss of generality, we assume that the Internet arriving data follows Pareto distribution and new users initiate authentication process when the UE is on the

move. In the Matlab simulation, SDN-enabled UE-specific SCI is pre-collected and transferred to relevant cells on the projected UE moving path, to realize fast verification. On the other hand, traditional authentication protocol establishes new authentications in each HetNet cell. Here we use two publicly available OpenFlow controllers' data as the representative to show the performance of SDN processing[69], i.e., NOX-MT and Beacon. NOX-MT is a multi-threaded successor of NOX, while Beacon is a Java controller built by David Erickson at Stanford [21].

Fig. 6.7 shows the comparison of authentication delay versus network utilization rates. Here network utilization is defined as the ratio of total data arrival rate and the controller processing rate. Here network utilization rate is used as a reflection of the various load situation of the network in order to provide more accurate latency performance. It can be seen from Fig. 6.7 that when the network load is fairly low, authentication delay is not a problem for both SDN and non-SDN networks. With more user arrivals and increased network load, SDN-enabled fast authentication still keeps the latency under 1ms most of the time, which meets the 5G latency requirement. NOX-MT and Beacon-enabled solutions perform 30% and 14.29% better than traditional handover authentication protocol in latency reduction with the commonly

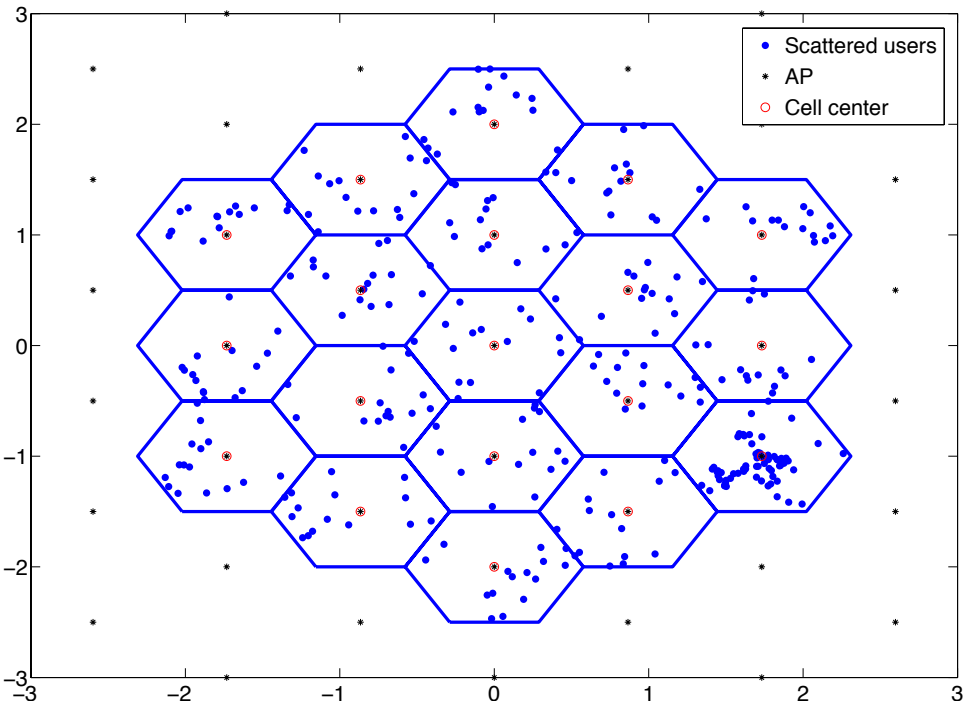

Figure 6.6: Simulation layout of 5G small cells with proportional axis (1 = 300m)

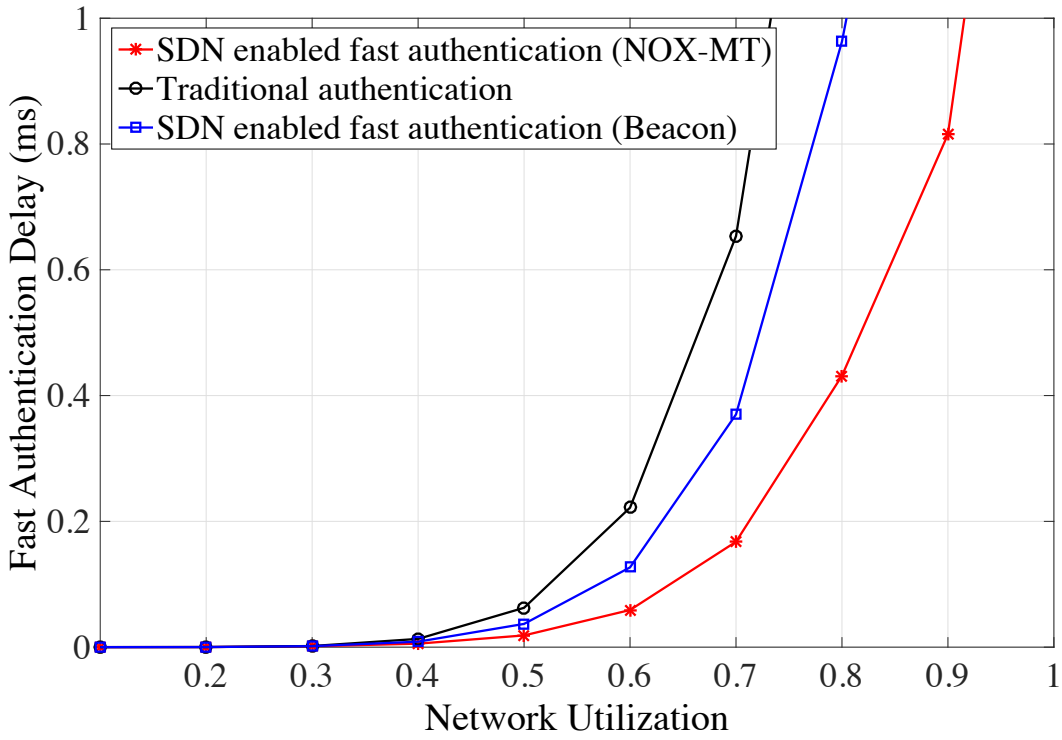

Figure 6.7: Simulation results of SDN-enabled fast authentication delay compared with traditional cryptographic authentication method.

used deployment of an eight-core machine, 2GHz CPUs and 32 switches in [69]. It is evident that SDN-enabled fast authentication has better performance in meeting the critical latency requirement in 5G while maintaining the SDN flexibility, programmability for 5G networks.

6.6.2.2 Secure level

To measure the secure level of the proposed weighted SCI based fast authentication algorithm, we conduct the Matlab simulation, which focuses on the probability of detection, P_D, regarding the required P_{FA}. A weighted combination of N attributes are used for authentication, and then likelihood ratio test is applied to make the authentication decision. In the simulation, the weight value is set to be equal to emphasize the difference in the different number of SCI combination. However, in the realistic scenario, it can be decided according to the difficulties of spoofing or optimized for the purpose of maximizing the detection probability. In each simulation, different SCI attributes number is set, i.e., $N = 1, 3, 5$, and the likelihood test offset is defined to be within 2%.

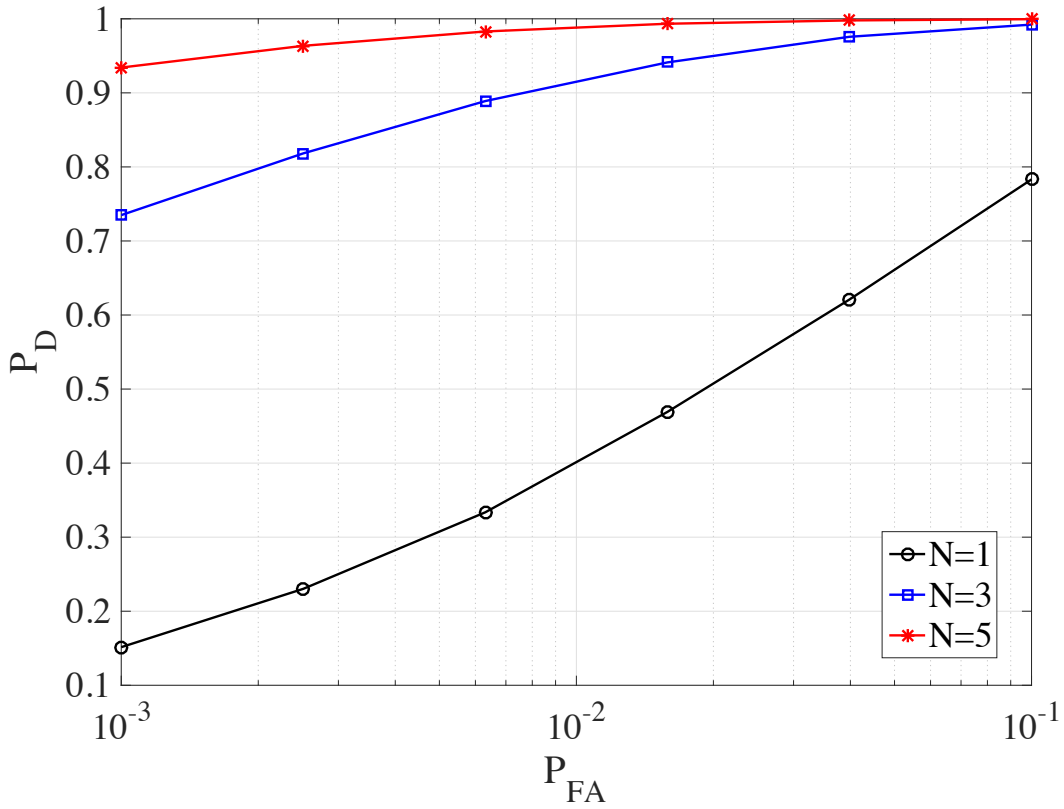

Figure 6.8: Weighted SCI based fast authentication algorithm performance with different number of attributes N

The simulation results are shown in Fig. 6.8. It is clear that the detection probabilities of using more than one attributes (i.e., the red and blue curves) are significantly improved compared with just using single characteristic (i.e., black curve). The P_D of both $N = 3$ and $N = 5$ remain above 70% in this simulation, and continue to rise to close to 100% when P_{FA} reaches 4%. Moreover, the weight of different attributes can be adjusted according to the actual application requirements.

The performance of the authentication algorithms under a different number of observations M is also presented in Fig. 6.9. We can see that under all scenarios of $N = 1$, $N = 3$ and $N = 5$, increasing the number of observations helps to boost the probabilities of detection, however, of cause, at the cost of complexity and latency due to the data collection procedure. It is also evident from the figure that under the same number of observations $M = 20$ (i.e., the blue dotted line and the solid line), which is a reasonable number of observations, multiple SCI increased P_D by 79% (from 15% to 74%). Therefore, it is safe to draw the conclusion that

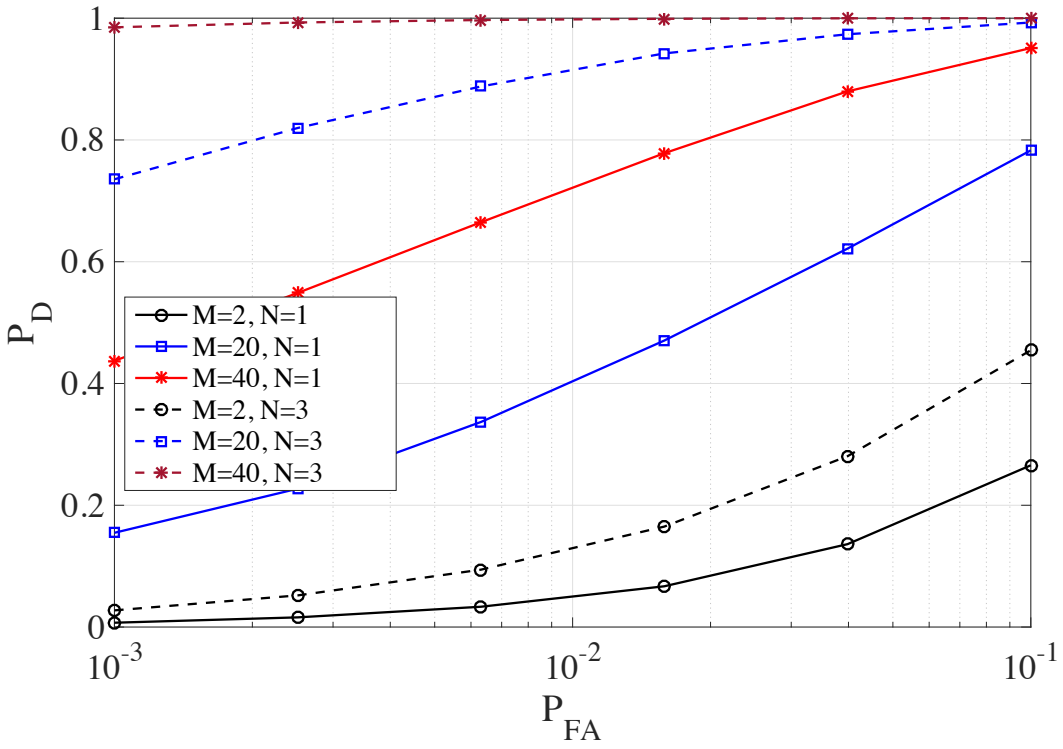

Figure 6.9: Weighted SCI based fast authentication algorithm performance with different number of observations M.

multiple SCI combination improves the detection probability while decrease authentication latency at the same time.

6.7 Chapter Summary

The emerging 5G communications bring new challenges on existing security provisioning mechanism, especially on latency and complexity. In this chapter, we introduce SDN into 5G networks and propose a fast authentication method based on weighted security context transfer. The new algorithm requires no changes to the existing UE and AP hardware, while significantly simplifies the authentication procedure and reduces handover latency through non-cryptographic technique. Based on the proposed secure context transfer, security in SDN-enabled 5G networks becomes a monitored seamless procedure. Additionally, SDN-enabled security framework also possesses high levels of tolerance to network failures with the pre-shared secure context information. Operators can even choose to switch off/on low loaded

cells if the users approaching these cells are not going to exceed a certain threshold according to the SCI information to save more energy. The proposed SDN-enabled non-cryptographic security technique provides effective solutions in addressing several challenges in 5G HetNet.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

5th generation mobile networks (5G), also referred to as beyond 2020 mobile communications systems, represent the next major phase of the mobile telecom industry, going beyond the current Long Term Evolution (LTE) and IMT-advanced systems. In addition to increased peak bit rates, better coverage, and higher spectrum spectral efficiency, 5G systems are required to enable ultra-reliable and low-latency communications (URLLC), support Internet of Things with potential numbers of diverse connectable devices, including massive machine type communications (mMTC) devices. All in all, 5G should be cost-efficient, flexible deployable, elastic, and above all programmable. To realize the flexible and elastic deployments and cope with the ever growing mobile data traffic, lots of research has been conducted in terms of 5G network management, resource allocation, as well as security provisioning. However, technical challenges still exist due to the complicated communication scenario. This thesis carries out a comprehensive study on 5G techniques which mainly explored the efficient management of heterogeneous networks: the resource management solutions including traffic offloading and spectrum sharing in high load scenario and 5G-VANET, and the security management in heterogeneous small cell deployment. We developed a number of effective and efficient schemes for these problems through the coordination of SDN platform.

The contributions that have been made in this thesis and the conclusions drawn from these contributions can be summarized as follows:

In Chapter 3, the novel SDN-based Wi-Fi data offloading and load balancing algorithms were proposed to cope with the increased data traffic and heterogeneous network structure. The new algorithms utilized the controller's global view of the network to take more informed decisions for efficient resource management. We also analyzed the performance of the proposed algorithms under realistic load conditions. To this end, we first introduced a queuing model with Pareto arrivals to investigate the processing and forwarding delays incurred due to the SDN architecture. Then, we analyzed the performance of the proposed SDN-based partial data offloading scheme in terms of the threshold miss probability and the amount of data offloaded successfully onto Wi-Fi. Through simulations, it was shown that partial data offloading saves primary resources and decreases threshold miss probability by $20\% \sim 50\%$, which ultimately improves the application performance at the user end. Furthermore, the simulation results also confirmed that SDN-based LB outperforms the baseline methods by minimizing the number of required handovers by 50% and by balancing the loads more evenly across multiple cells. Our results and discussions showed that the delay incurred by SDN is well within the acceptable limits for most applications. Particularly, it has been demonstrated that SDN-based solutions perform better for large data traffic with high delay tolerance. All in all, SDN is proved to be a suitable enabling technology for introducing intelligence within the wireless networks and for providing fine-grained control to the network operators.

To evaluate the resource management schemes in a particular network, the traffic offloading performance in 5G-VANET was studied in Chapter 4. In this chapter, we propose to integrate SDN into 5G-VANET and thus provide a programmable platform in addressing the challenges of dynamic vehicle communications. With the anticipated arrival of self-driving vehicles and dramatic growth of in-vehicle mobile data traffic, supporting of dynamic vehicle communications in 5G HetNets are expected to be extremely challenging, due to fast varying network topology and high complexity of the heterogeneous infrastructure. Through the proposed SDN-enabled adaptive vehicle clustering and dual cluster head scheme, signaling overhead of VANET is significantly reduced along with improved communication quality. The proposed cluster head selection also guarantees the seamless access to the operators' services for the cluster users. To accommodate the varying traffic over the trunk link and reduce the latency during traffic distribution, adaptive trunk link transmission scheme and cooperative

communication of mobile gateway candidates were proposed for the aggregated V2I traffic transmission in this integrated network. Simulation results show that SDN coordinated vehicle clustering and beamformed transmission are suitable to support fast varying traffic conditions with enormous dynamic range.

An orchestrated spectrum sharing architecture that integrates the distributed systems into an amalgamated network with real-time information exchange and a 3D interference map based spectrum sharing algorithm were proposed in Chapter 5. In order to cope with spectrum scarcity challenges, the limited sensing capability of devices and lack of timely information exchange between coexisting heterogeneous networks (HetNets), an orchestrated spectrum sharing approach that integrates the distributed located users, base stations (BS), incumbent stations, and the Software-Defined Networking (SDN) controller into an amalgamated network with real-time information exchange was proposed in this Chapter. To adequately protect incumbent users and efficiently share the pooled spectrum resources, real-time 3D interference map was considered to guide the spectrum access based on the SDN global view.

The last contribution of this thesis was the reduction of the latency during user authentication procedure using the non-cryptographic method and simplified privacy protection. In Chapter 6, we introduce SDN into 5G networks and propose a fast authentication method based on weighted security context transfer. The new algorithm requires no changes to the existing UE and AP hardware, while significantly simplifies the authentication procedure and reduces handover latency through non-cryptographic technique. Based on the proposed secure context transfer, security in SDN-enabled 5G networks becomes a monitored seamless process. Additionally, SDN-enabled security framework also possesses high levels of tolerance to network failures with the pre-shared secure context information. Operators can even choose to switch off/on low loaded cells if the users approaching these cells are not going to exceed a certain threshold according to the SCI information to save more energy. The proposed SDN-enabled non-cryptographic security technique provides practical solutions in addressing several challenges in 5G HetNet.

In summary, this thesis proposed several solutions to the urgent challenges in 5G HetNets using SDN programmable platform so that the network performance could be improved. With more efforts on user QoS and application specific performance enhancement, next generation

5G communications are believed to be adaptable to user demand and thus improve everyday life.

7.2 Future Work

The contributions presented in this dissertation for SDN-enabled 5G HetNets can be extended or used to explore new research topics. In the future, some aspects of the proposed algorithms are also worthwhile to be further investigated. Some potential research works are summarized as follows:

- The partial mobile data offloading with load balancing design in Chapter 3 can reduce cellular network burden, while taking the latency requirement of user application into account. Under the condition that user application has stringent delay requirement, e.g., live video, the data traffic will be transmitted over cellular and Wi-Fi networks at the same time, and the traffic amount to be transmitted in each network is calculated according to the proposed algorithm. However, the delay incurred due to the traffic division before transmission and the traffic aggregation afterward is still to be studied. Moreover, the additional signaling or overhead because of the simultaneous transmission also need to be considered. How to optimally make the data offloading decision and the partial transmission could be further investigated, especially combined with latest releases of Wi-Fi and cellular standard and protocols.

- The optimization problem of cluster head selection in Chapter 4 takes the battery level and channel condition of the mobile devices into consideration. The toolbox was used to obtain the results of the optimization. However, it is complex and takes more time. A new study could be started on solving the optimization problem mathematically and reduce the running time of the algorithm.

- The non-cryptographic based fast authentication scheme in Chapter 6 utilizes multiple user attributes as secure context information and simplifies authentication procedure, especially in 5G small cells. The performance of the proposed algorithm was analyzed in

terms of the detection probability compared with single attribute authentication. However, the comparison between non-cryptographic authentication method and traditional cryptographic schemes are also necessary to verify the performance. The indicator of the secure level that can accommodate both authentication schemes are to be studied, and how to optimally select the authentication scheme for a particular type of communication procedure or network could be further investigated.

- All the proposed algorithms in this thesis were simulated in MATLAB. In other words, although we referred the data of SDN processing capability from published papers and testbed experiments from other authors, the performance evaluations of the proposed algorithms were based on the simulated tests. Since applicability is important in SDN, practical implementation and experiments could be conducted to evaluate the algorithms in the future.

Bibliography

[1] Wi-fi. en.wikipedia.org/wiki/Wi-Fi.

[2] L. Chen, J. Ji, and Z. Zhang. Wireless security: Models, threats, and solutions. *Higher Education Press, Heidelberg*, 2013.

[3] K. Lee, J. Lee, and Y. Yi. Mobile Data Offloading : How Much Can WiFi Deliver? *IEEE/ACM Transactions on networking*, 21(2):536–551, 2013.

[4] H. Elsawy, H. Dahrouj, T. Y. Al-naffouri, and M. s. Alouini. Virtualized cognitive network architecture for 5g cellular networks. *IEEE Communications Magazine*, 53(7):78–85, July 2015.

[5] G. Wang, G. Feng, S. Qin, and R. Wen. Efficient traffic engineering for 5g core and backhaul networks. *Journal of Communications and Networks*, 19(1):80–92, February 2017.

[6] Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update, 2016-2021. www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2017.

[7] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang. Enhanced intercell interference coordination challenges in heterogeneous networks. *IEEE Wireless Communications*, 18(3):22–30, June 2011.

[8] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, and J. Lu. 5G on the Horizon: Key Challenges for the Radio-Access Network. *IEEE Vehicular Technology Magazine*, 8(3):47–53, September 2013.

[9] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano A. C. K. Soong, and J. C. Zhang. What will 5g be? *IEEE Jour. Sel. Areas Commun.*, 32(6):1065–1082, June 2014.

[10] A. M. Akhtar, X. Wang, and L. Hanzo. Synergistic Spectrum Sharing in 5G HetNets: A Harmonized SDN-enabled Approach. *IEEE Commun. Mag.*, 54(1):40–47, January 2016.

[11] T. Taleb, A. Ksentini, and R. Jantti. "anything as a service" for 5g mobile systems. *IEEE Network*, 30(6):84–91, November 2016.

[12] A detailed look at hotspot 2.0. www.ruckuswireless.com/technology/hotspot2, Sept 2015.

[13] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen. From lte to 5g for connected mobility. *IEEE Communications Magazine*, 55(3):156–162, March 2017.

[14] B. Jane et al. Effects of next generation vehicles on travel demand and highway capacity. *FP Think Working Group*, pages 10–11, January 2014.

[15] IEEE Standard for Information TechnologyLocal and Metropolitan Area NetworksSpecific RequirementsPart 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments. Ieee standard 802.11p-2010, Jul.15 2010.

[16] Intelligent transportation systems committee (vt/its). http://standards.ieee.org/develop/wg/1609WG.html, 2016.

[17] Y. Chen, M. Fang, S. Shi, W. Guo, and X. Zheng. Distributed multihop clustering algorithm for vanets based on neighborhood follow. *EURASIP J. Wireless Commun. Netw.*, 1:1–12, April 2015.

[18] C. Cooper, D. Franklin, M. Ros, F. Safaei, and M. Abolhasan. A comparative survey of vanet clustering techniques. *IEEE Communications Surveys Tutorials*, 19(1):657–681, Firstquarter 2017.

[19] G. Papastergiou, G. Fairhurst, D. Ros, A. Brunstrom, K. J. Grinnemo, P. Hurtig, N. Khademi, M. Txen, M. Welzl, D. Damjanovic, and S. Mangiante. De-ossifying the internet transport layer: A survey and future perspectives. *IEEE Communications Surveys Tutorials*, 19(1):619–639, Firstquarter 2017.

[20] Open network foundation (onf). software defined networking: the new form for networks [eb/ol]. https://www.opennetworking.org/images/stories/downloads/white- papers/wp-sdn-newnorm.pdf.

[21] A.N. Bruno, M. Marc, and N. Xuan-nam. A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks. *IEEE Communications Surveys and Tutorials*, 16(3):1617–1634, February 2014.

[22] M. Nick, A. Tom, and B. Hari. Openflow: Enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review*, pages 69–74, 2008.

[23] A. T. Campbell, I. Katzela, K. Miki, and J. Vicente. Open signaling for atm internet and mobile networks (opensig'98). *ACM SIGCOMM Computer Communication Review*, 29(1), 1999.

[24] D.L. Tennenhouse, J.M. Smith, W.D. Sincoskie, D.J. Wetherall, and G.J. Minden. A survey of active network research. *IEEE Commun. Mag.*, 35(1), 1997.

[25] Devolved control of atm networks [online]. Available: http://www.cl.cam.ac.uk/research/srg/netos/old-projects/dcan/#pub.

[26] A. Greenberg, G. Hjalmtysson, D. Maltz, A. Myers, J. Rexford, G. Xie, H. Yan, J. Zhan, and H. Zhang. A clean slate 4d approach to network control and management. *ACM SIGCOMM Computer Communication Review*, 35(5), 2005.

[27] Open flow switch specification [online]. `www.opennetworking.org`, February 2011.

[28] K-K. Yap, R. Sherwood, M. Kobayashi, T-Y. Huang, M. Chan, N. Handigol, N. McKeown, and G. Parulkar. Blueprint for introducing innovation into wireless mobile networks. *Proceedings of ACM SIGCOMM workshop*, pages 25–32, 2010.

[29] L. E. Li, Z. M. Mao, and J. Rexford. Toward software-defined cellular networks. In *2012 European Workshop on Software Defined Networking*, pages 7–12, Oct 2012.

[30] L. Vanbever X. Jin, L. E. Li and J. Rexford. Softcell: Scalable and flexible cellular core network architecture. *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pages 163–174, 2013.

[31] K. Pentikousis, Y. Wang, and W. Hu. Mobileflow: Toward software-defined mobile networks. *IEEE Communications Magazine*, 51(7):44–53, July 2013.

[32] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann. Applying NFV and SDN to LTE Mobile Core Gateways, the Functions Placement Problem. In *Proceedings of the 4th Workshop on All Things Cellular: Operations, Applications, & Challenges*, AllThingsCellular '14, pages 33–38, New York, NY, USA, 2014. ACM.

[33] S. Tomovic, M. Pejanovic-Djurisic, and I. Radusinovic. SDN based mobile networks: Concepts and benefits. *Wireless Personal Communications*, 78(3):16291644, 2014.

[34] A. Gudipati, D. Perry, L. E. Li, and S. Katti. SoftRAN: Software defined radio access network. *2th ACM SIGCOMM workshop on Hot topics in software defined networking*, page 2530, 2013.

[35] B. Soret and K. I. Pedersen. Centralized and distributed solutions for fast muting adaptation in lte-advanced hetnets. *IEEE Transactions on Vehicular Technology*, 64(1):147–158, Jan 2015.

[36] A Aijaz, H. Aghvami, and M. Amani. A survey on mobile data offloading: technical and business perspectives. *IEEE Wireless Communications*, 20(2):104–112, April 2013.

[37] Mobile broadband access at home, August 2008.

[38] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting Mobile 3G Using WiFi. *ACM MobiSys*, pages 209–222, June 2010.

[39] S-I. Sou. Mobile Data Offloading With Policy and Charging Control in 3GPP Core Network. *IEEE Transactions on Vehicular Technology*, 62(7):3481–3486, September 2013.

[40] G. Lin, I. George, H. Jianwei, and T. Leandros. Economics of Mobile Data Offloading. *IEEE INFOCOM Workshop*, pages 3303–3308, 2013.

[41] J. Lee, Y. Yi, S. Chong, and Y. Jin. Economics of WiFi Offloading: Trading Delay for Cellular Capacity. *IEEE INFOCOM*, pages 357–362, 2013.

[42] Study of heterogeneous networks management. TS 32.835 Release 12, 3GPP, March 2013.

[43] T. Tarik, B. Abderrahim, and K.B. Letaif. Towards an effective risk-conscious and collaborative vehicular collision avoidance systems. *IEEE Trans. Veh. Technol.*, 59(3):1474–1486, March 2010.

[44] B. Abderrahim, T. Tarik, and S. Rajarajan. Dynamic clustering-based adaptive mobile gateway management in integrated VANET-3G Heterogeneous Wireless Networks. *IEEE Journal on Selected Areas in Communications*, 29(3):559–570, March 2011.

[45] R. Kwan, R. Arnott, R. Paterson, R. Trivisonno, and M. Kubota. On Mobility Load Balancing for LTE Systems. *IEEE Vehicular Technology Conference Fall*, 1(5):6–9, September 2010.

[46] J. Haydar, A. Ibrahim, and G. Pujolle. A New Access Selection Strategy in Heterogeneous Wireless Networks Based on Traffic Distribution. *Wireless Days 1st IFIP*, 1(5):24–27, November 2008.

[47] D. Lopez-Perez, I. Guvenc, and Xiaoli Chu. Theoretical analysis of handover failure and ping-pong rates for heterogeneous networks. In *IEEE International Conference on Communications (ICC)*, pages 6774–6779, June 2012.

[48] Technical specification group service and system aspects; 3gpp system architecture evolution (sae); security architecture (rel 11). TS 33.401 V11.5.0, 3GPP, 2012.

[49] D. He, C. Chen, S. Chan, and J. Bu. Secure and efficient handover authentication based on bilinear pairing functions. *IEEE Transactions on wireless communications*, 11(1):48–53, January 2012.

[50] J. Choi and S. Jung. A handover authentication using credentials based on chameleon hashing. *IEEE communication letters*, 14(1):54–56, January 2010.

[51] C. Jin, L. Hui, M. Maode, Z. Yueyu, and L. Chengzhe. A Simple and Robust Handover Authentication between HeNB and eNB in LTE Networks. *Computer Networks*, 56(8):2119–2131, May 2012.

[52] L. Cai, S. Machiraju, and H. Chen. Capauth: a capability-based handover scheme. *IEEE Infocom*, pages 1–5, 2010.

[53] D. He, C. Chen, J. Bu, S. Chan, and Y. Zhang. Security and efficiency in roaming services for wireless networks: Challenges, approaches, and prospects. *IEEE communication magazine*, 51(2):142–150, February 2013.

[54] K. Zeng, K. Govindan, and P. Mohapatra. Non-cryptographic authentication and identification in wireless networks. *IEEE wireless communication*, 17(5):56–62, October 2010.

[55] A. M. Hatami, M. Mirmohseni, and F. Ashtiani. A new data offloading algorithm by considering interactive preferences. In *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 1–6, Sept 2016.

[56] K-K. Yap, M. Kobayashi, R. Sherwood, T-Y. Huang, M. Chan, N. Handigol, and N. McKeown. OpenRoads: Empowering research in mobile networks. *ACM SIGCOMM Computer Communication Review*, 40(1):125–126, 2010.

[57] P. Dely, A. Kassler, L. Chow, N. Bambos, N. Bayer, H. Einsiedler, C. Peylo, D. Mellado, and M. Sanchez. A software-defined networking approach for handover management with real-time video in WLANs, June 2013.

[58] Technical specification group services and system aspects; policy and charging control architecture. TS 23.203 Ver. 9.5.0, 3GPP, June 2010.

[59] C-H. Ko and H-Y. Wei. On-Demand Resource-Sharing Mechanism Design in Two-Tier OFDMA Femtocell Networks. *IEEE Transactions on Vehicular Technology*, 60(3):1059–1071, March 2011.

[60] Z. Lin, L. Yu, Z. Mengru, J. Shucong, and D. Xiaoyu. A Two-Layer Mobility Load Balancing in LTE self-organization networks. *IEEE International Conference on Communication Technology (ICCT)*, pages 925–929, September 2011.

[61] M. Jarschel, S. Oechsner, and D. Schlosser. Modeling and performance evaluation of an openflow architecture. *Proceedings of the 2011 23rd International Teletraffic Congress*, pages 1–7, 2011.

[62] A.C. Feldmann, A. Gilbert and W. Willinger. Data networks as cascades: Investigating the multifractal nature of the internet. *ACM SIGCOMM Computer Communication Review*, 28:42–55, 1998.

[63] K.S. Munasinghe and A Jamalipour. Evaluation of session handoffs in a heterogeneous mobile network for pareto based packet arrivals. *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, April 2009.

[64] M.E. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic : Evidence and Possible Causes. *IEEE/ACM Transactions on networking*, 5(6):835–846, December 1997.

[65] G.R. Dattatreya. *Performance Analysis of Queuing and Computer Networks*. Chapman Hall/CRC Computer and Information Science Series. Taylor Francis, 2008.

[66] M. Jarschel. *An Assessment of Applications and Performance Analysis of Software Defined Networking*. University Wurzburg, 2014.

[67] X. Li, H. Lu, and H. Lu. QoS Analysis of Self-Similar Multimedia Traffic with Variable Packet Size in Wireless Networks. *IEEE Vehicular Technology Conference Fall*, 1(5):2–5, September 2013.

[68] Robert G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2014.

[69] T. Amin, G. Sergey, and G. Yashar. On controller performance in software-defined networks. *Proc. HotICE*, 2012.

[70] P. Jonathan, S. Florian, and K. Frank. Pseudonym Schemes in Vehicular Networks: A Survey. *IEEE Commun. Surveys & Tutorials*, 17(1):228–255, 2015.

[71] H. Li, M. Dong, and K. Ota. Control Plane Optimization in Software-Defined Vehicular Ad Hoc Networks. *IEEE Trans. on Vehicular Technology*, 65(10):7895–7904, October 2016.

[72] X. Duan and X. Wang. Authentication handover and privacy protection in 5g hetnets using software-defined networking. *IEEE Communications Magazine*, 53(4):28–35, April 2015.

[73] I. Christoph, D. Bjoern, P. Markus, and W. Christian. Channel sensitive transmission scheme for v2i-based floating car data collection via lte. *IEEE International Conference on Communication*, pages 88–92, 2012.

[74] X.Duan, X.Wang, and Y. Liu. SDN Enabled Dual Cluster Head Selection and Adaptive Clustering in 5G-VANET. *IEEE VTC Fall*, 2016.

[75] S. Jia, S. Hao, X. Gu, and L. Zhang. Analyzing and Relieving the Impact of FCD Traffic in LTE-VANET Heterogeneous Network. *International Conference on Telecommunications (ICT)*, pages 88–92, 2014.

[76] J. S. Kaufmann. Blocking in a shared resource environment. *IEEE Transactions on communications*, 29(10):1474–1481, October 1981.

[77] H. Zaaraoui, Z. Altman, and E. Altman. Beam Focusing Antenna Array Technology for Non-Stationary Mobility. *IEEE Wireless Commun. and Networking Conference*, pages 1–6, April 2016.

[78] S. Zhang et al. Sparse Code Multiple Access: An Energy Efficient Uplink Approach for 5G Wireless Systems. *IEEE GLOBECOM*, pages 4782–4787, 2014.

[79] X. Duan, A. M. Akhtar, and X. Wang. Software-defined networking-based resource management: data offloading with load balancing in 5g hetnet. *EURASIP Journal on Wireless Communications and Networking*, 181, Junel 2015.

[80] D. B. Rawat, D. C. Popescu, G. Yan, and S. Olariu. Enhancing vanet performance by joint adaptation of transmission power and contention window size. *IEEE Transactions on Parallel and Distributed Systems*, 22(9):1528–1535, Sept 2011.

[81] E. A. Feukeu, S. M. Ngwira, and T. Zuva. Doppler shift signature for bpsk in a vehicular network: Ieee 802.11p. In *2015 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1744–1749, Aug 2015.

[82] G. Song and J. Cheng. Distance Enumerator Analysis for Multi-User Codes. *IEEE ISIT*, pages 3137–3141, 2014.

[83] Y. Liu, X. Wang, and X. Duan. Aggregated V2I Communications for Improved Energy Efficiency using Non-Orthogonal Multiplexed Modulation. *IEEE VTC Fall*, 2016.

[84] M. Mueck, W. Jiang, G. Sun, H. Cao, E. Dutkiewicz, and S. Choi. Novel Spectrum Usage Paradigms for 5G. *IEEE Cognitive Networks*, November 2014.

[85] S. I. Park, W. Li, J. Y. Lee, Y. Wu, X. Wang, S. Kwon, B. M. Lim, H. M. Kim, N. Hur, L. Zhang, and J. Kim. Atsc 3.0 transmitter identification signals and applications. *IEEE Transactions on Broadcasting*, 63(1):240–249, March 2017.

[86] J. Khun-Jush, P. Bender, B. Deschamps, and M. Gundlach. Licensed shared access as complementary approach to meet spectrum demands: Benefits for next generation cellular systems. *Proc. ETSI Workshop Reconfigurable Radio Syst.*, pages 1–7, 2013.

[87] M. Matinmikko, H. Okkonen, M. Palola, S. Yrjola, P. Ahokangas, and M. Mustonen. Spectrum sharing using licensed shared access: the concept and its workflow for lte-advanced networks. *IEEE Wireless Communications*, 21(2):72–79, April 2014.

[88] B. Yu, L. Yang, and H. Ishii. Load balancing with 3-d beamforming in macro-assisted small cell architecture. *IEEE Transactions on Wireless Communications*, 15(8):5626–5636, Aug 2016.

[89] L. Bernado, N. Czink, T. Zemen, and P. Belanovic. Physical layer simulation results for ieee 802.11p using vehicular non-stationary channel model. In *2010 IEEE International Conference on Communications Workshops*, pages 1–5, May 2010.

[90] J. B. Andersen, T. S. Rappaport, and S. Yoshida. Propagation measurements and models for wireless communications channels. *IEEE Communications Magazine*, 33(1):42–49, Jan 1995.

[91] 3.5 ghz band / citizens broadband radio service. `https://www.fcc.gov/rulemaking/12-354`, July 2016.

[92] F.J. Liu, X. Wang, and S. Primak. A two dimensional quantization algorithm for cir-based physical layer authentication. *IEEE International Conference on Communications*, pages 4724–4728, June 2013.

[93] K. Zeng, K. Govindan, and P. Mohapatra. Non-cryptographic authentication and identification in wireless networks. *IEEE Wireless Communications*, 17(5):56–62, October 2010.

[94] P. Hao, X. Wang, and A. Behnad. Performance enhancement of i/q imbalance based wireless device authentication through collaboration of multiple receivers. *IEEE International Conference on Communications*, pages 939–944, June 2014.

[95] W. Hou, X. Wang, J. Y. Chouinard, and A. Refaey. Physical layer authentication for mobile systems with time-varying carrier frequency offsets. *IEEE Transactions on Communications*, 62(5):1658–1667, May 2014.

[96] D. Son; A. Helmy; B. Krishnamachari. The effect of mobility-induced location errors on geographic routing in mobile ad hoc sensor networks: analysis and improvement using mobility prediction. *IEEE Transactions on Mobile Computing*, 3(3):233–245, 2004.

[97] V. Brik, S. Banerjee, M. Gruteser, and S. Oh. Wireless device identification with radiometric signatures. *ACM International Conference on Mobile Computing and Networking*, pages 116–127, 2008.

[98] S.M. Kay. Fundamentals of statistical signal processing: Detection theory. *Prentice Hall Signal Processing Series*, 1998.

[99] Priority queueing (nonpreemptive). http://www.ece.virginia.edu/mv/edu/715/lectures/PQ.pdf.

Curriculum Vitae

Name:	Xiaoyu Duan
Post-Secondary Education and Degrees:	2013 - present, PhD Electrical and Computer Engineering The University of Western Ontario London, Ontario, Canada 2010 - 2013, M.E.Sc Signal and Information Processing Beijing University of Posts and Telecommunications Beijing, China 2006 - 2010, B.Eng (Hons) Communication Engineering Tianjin University Tianjin, China
Honours and Awards:	NSERC CREATE Scholar, 2013-2017
Related Work Experience:	Teaching Assistant The University of Western Ontario 2013 - 2017 Research Assistant The University of Western Ontario 2013-2017

Publications:

[1] X.Duan and X.Wang, "Partial Mobile Data Offloading with Load Balancing in Heterogeneous Cellular Networks Using Software-Defined Networking," in *Proc. IEEE PIMRC*, September 2014.
[2] X.Duan, A.Akhtar and X.Wang, "Software-Defined Networking based Resource Management: Data Offloading with Load Balancing in 5G HetNet," in *EURASIP Journal on Wireless*

Communications and Networking, June 2015.
[3] X.Duan and X.Wang, "Authentication Handover and Privacy Protection in 5G HetNet Using Software-Defined Networking," in *IEEE Communications Magazine*, April 2015.
[4] X.Duan and X.Wang, "Fast authentication in 5G HetNet through SDN enabled Weighted Secure-Context-Information transfer," in *IEEE ICC*, May 2016.
[5] X.Duan, X.Wang, Y. Liu and K. Zheng, "SDN Enabled Dual Cluster Head Selection and Adaptive Clustering in 5G-VANET," in *IEEE VTC Fall*, September 2016.
[6] X.Duan, Y. Liu, X.Wang, "SDN Enabled 5G-VANET: Adaptive Vehicle Clustering and Beamformed Transmission for Aggregated Traffic," in *IEEE Communications Magazine*, July 2017.
[7] Y. Liu, X.Wang, X.Duan, and H. Lin, "Aggregated V2I Communications for Improved Energy Efficiency using Non-Orthogonal Multiplexed Modulation," in *IEEE VTC Fall*, September 2016.
[8] Y. Liu, X.Duan, G. Boudreau, A. B. Sediq, and X.Wang, "Adaptive Beamforming based Inband Fronthaul for Cost-Effective Virtual Small Cell in 5G Networks," in *IEEE Global Communications Conference*, accepted.